W203 Lab 3 - COVID-19

August 3, 2020

Blake Allen, Sam Shih, Mike Fan

Summer 2020 Section 1, Oleg Melnikov

1 Introduction

The SARS-Coronavirus disease 2019 or commonly known as Covid-19 has rocked the world into a global pandemic of unprecedented proportion. In this project we will focus on identifying which government policies have been the most effective at minimizing Covid-19 infection rate. We also measure policy effectiveness by assessing the length or duration of policies put into effect and its corresponding impact on infection rate.

2 Exploratory Data Analysis (EDA)

2.1 Imputation of Missing Values

We noticed certain states had 1 or more missing values for some variables and did our best to fill in missing gaps based on what we believe to be trusted sources of information. The following states closure of non-essential business (source: COVID-19 US state policy database 07_07_2020)

- **Arkansas**
- **Florida**
- **Georgia**
- **Nebraska**
- **North Dakota**
- **Oklahoma**
- **Virginia**
- **Wyoming**

Additional imputation by state below:

- **Connecticut**
  - Shelter-in-place order: 3/20/2020 (source: Executive Order 7H)
  - End/relaxation of shelter-in-place: 5/20/2020 (source: Executive Order No. 7X)
- **Hawaii**
  - End/relaxation of shelter-in-place: 7/31/2020 (source: Ninth Supplementary Proclamation Related to the Covid-19 Emergency)
- **Kentucky**
  - Total test results: 395,647 as of 7/5/2020. We determined that the test results data were compiled on July 5, 2020 by corroborating totalTestResults with Alabama's historical

data that's tracked and made available by The COVID Tracking Project. (source: The COVID Tracking Project)
  - Start of shelter-in-place: 3/22/2020 (source: Executive Order 2020-246)
- **New Jersey**
  - Start of shelter-in-place: 3/21/2020 (source: Executive Order No. 107)
  - End/relaxation of shelter-in-place: 6/9/2020 (source: Executive Order No. 108)
- **South Dakota**
  - Start of shelter-in-place: 3/23/2020 (source: Executive Order 2020-08)
  - End/relaxation of shelter-in-place: 4/28/2020 (source: Executive Order 2020-20)
- **Texas**
  - Start of shelter-in-place: 3/31/2020 (source: Executive Order by the Governor 22)
  - End/relaxation of shelter-in-place: 4/30/2020 (source: Executive Order by the Governor 23)
- **Utah**
  - Start of shelter-in-place: 3/27/2020 (source: "Stay Home, Stay Safe" Directive)
  - End/relaxation of shelter-in-place: 5/1/2020 (source: Executive Order No.19)
  - Closure of non-essential businesses: 3/19/2020; restaurant dine-in were ordered to shutdown on 3/19 and the variable "Began to reopen businesses statewide" was based on the reopen date of dine-in restaurants (source: COVID-19 US state policy database 07_07_2020)
- **Ohio**
  - Children 0-18: 22%
  - Adults 26-34: 13%
  - Adults 35-54: 14%
  - The above demographic values for Ohio are approximations based off of census data (source: United States Census Bureau)

2.2 Data Enrichment

In addition to the provided dataset, we also added and derived new variables in Excel to further support the analysis of our research question:

- **party2016:**Winning political party to the 2016 Presidential Election for each state. We are interested in understanding if political allegiance had any significant impact to the timeliness of legislative action and its impact to infection rate (source: MIT Election Data and Science Lab)
- **party2016_votepercent:** Percentage of total popular vote the winning party received (source: MIT Election Data and Science Lab)
- **CasesInLast7Days_100k:** Since each state's population can vary widely, we control for population by calculating the number of positive cases in the last 7 days per 100,000 residents
- **totalTestResults_100k:** Since each state's population can vary widely, we control for population by calculating the number of test results in the last 7 days per 100,000 residents
- **infection rate:** Total number of positive cases divided by the total number of test results. This is also our dependent variable.
- **soe_to_sip:** Number of days between the state of emergency was declared and shelter-in-place was ordered. We're interested in understanding whether the timeliness of stay at home orders following the the state of emergency had any significant impact to fatality rate
  - For the state of California, Kentucky, New Mexico and New York, their respective shelter-

in-place order is still active. Thus, the value was derived by taking the difference between the shelter-in-place start date and 7/7/2020, which is the date of one of the sources used to compile the dataset (source: COVID-19 US state policy database 07_07_2020)

- **sip_binary:** Binary version of stay at home/shelter-in-place variable; coded a "1" for states that had a shelter-in-place order and a "0" otherwise
- **sip_start_end:** Measures the number of days the shelter-in-place order was active (difference between start and rescind date). For states that did not have a shelter-in-place order, the value is 0
- **soe_biz_close:** Represents the number of days between a state of emergency was declared and the shutdown of all non-essential businesses. A few select states did not have such policy, so the value is appropriately 0
- **biz_close_binary:** Binary version of closed non-essential businesses variable; coded a "1" for states that an explicit order from the governor's office to close all non-essential businesses and a "0" otherwise. Even if a state had shut down certain elements of non-essential businesses (i.e. dine-in restaurants, gyms, fashion retailer) but did not have an executive order from the state to shut down **all** non-essential business, we still coded it as 0
- **biz_close_open:** The number of days between the closure of non-essential businesses and business re-open date. There are 10 states in which had a statewide business re-open date, but did not have an initial closure date. This begs the question of how can a state re-open its businesses when it was never shutdown to close in the first place? Upon further examination
- **soe_face_mask:** Number of days between the state of emergency declaration and when face masks became mandatory for employees in public-facing businesses
- **face_mask_binary:** Binary version of mandate face mask use by employees in public-facing businesses variable; coded a "1" for states that put forth such order and a "0" otherwise
- **all_deathrate18:** Percent of each state's 2018 population that died for all causes

2.3 Data Cleansing

We begin our analysis by loading the data set and performing basic checks and inspections. Upon initial inspection, we observe there are 52 observations with 25 variables.

[ ]:
```
library(plyr)
library(dplyr)
library(ggplot2)
library(summarytools)
library(coin)
library(effsize)
library(tidyverse)
library(lsr)
library(corrplot)
library(rstatix)
library(BSDA)
library(cowplot)
library(ggrepel)
library(stargazer)
library(knitr)
library(kableExtra)
library(corrr)
```

```
library(ggcorrplot)
library(car)
library(lmtest)
library(sandwich)
```

[4]:
```
df_covid = read.csv("data/covid-19_dist0720_enriched.csv")
```

[5]:
```
df_covid2 = data.frame(df_covid)
```

[4]:
```
#Because variable names are case sensitive,
#we replace some of the more lengthy names with a shorter one for the sake of␣
 ↪parsimony and ease of parsing.
#Below is a mapping of the original names to its new value.
df_covid2 = df_covid2 %>% rename(state = State, CaseRate_100k = RatePer100000,␣
 ↪total_case = Total.Cases, total_death = Total.Death, soe = State.of.
 ↪emergency, sip_start_date = Stay.at.home..shelter.in.place, sip_end_date =␣
 ↪End.relax.stay.at.home.shelter.in.place,
                                 noness_bizclose_date = Closed.non.essential.
 ↪businesses, biz_reopen_date = Began.to.reopen.businesses.statewide,␣
 ↪fmask_emp_pfb_date = Mandate.face.mask.use.by.employees.in.public.facing.
 ↪businesses,
                                  wunemp_insure_max = Weekly.unemployment.
 ↪insurance.maximum.amount..dollars., pop_dens = Population.density.per.square.
 ↪miles, pop18 = Population.2018, perc_under_fed_povline18 = Percent.living.
 ↪under.the.federal.poverty.line..2018.,
                                  perc_atrisk_covid = Percent.at.risk.for.
 ↪serious.illness.due.to.COVID, all_death18 = All.cause.deaths.2018, age_0to18␣
 ↪= Children.0.18, age_19to25 = Adults.19.25, age_26to34 = Adults.26.34,␣
 ↪age_35to54 = Adults.35.54,
                                  age_55to64 = Adults.55.64, age_65plus = X65.)
```

[6]:
```
#Create a new variable that encodes Republican as a 0 and Democrat as a 1
#in order to operationalize party2016 for our analysis.
df_covid2$party16 = ifelse(df_covid2$party2016 == "Republican", 0, 1)
```

[60]:
```
#We also notice that party2016 and state are of character class type,
#so we will convert both into a factor.
#Additionally, all of the date-based variables
#(State of emergency, Stay at home/ shelter in place, end of shelter in place,
#business closure, business re-opening, face mask mandate of employees in␣
 ↪public-facing businesses)
#are of character class type. Thus, we will convert them into the date class.

df_covid2$party2016 = as.factor(df_covid2$party2016)
df_covid2$state = as.factor(df_covid2$state)
```

```
df_covid2$soe = as.Date(df_covid2$soe, format = "%m/%d/%Y")
df_covid2$sip_start_date = as.Date(df_covid2$sip_start_date, format ="%m/%d/%Y")
df_covid2$sip_end_date = as.Date(df_covid2$sip_end_date, format ="%m/%d/%Y")
df_covid2$noness_bizclose_date = as.Date(df_covid2$noness_bizclose_date, format␣
↪="%m/%d/%Y")
df_covid2$biz_reopen_date = as.Date(df_covid2$biz_reopen_date, format ="%m/%d/
↪%Y")
df_covid2$fmask_emp_pfb_date = as.Date(df_covid2$fmask_emp_pfb_date, format␣
↪="%m/%d/%Y")
```

[ ]:
```
#We also observed that there are 52 observations in the dataset
#while there are only 51 states in the United States.
#Upon closer inspection, the state of Arizona is entered twice,
#once with a capital A and the other with a lower case a.
#Additionally, the only 2 variables that showed different values between the␣
↪duplicate
#entries are total case and total death. When we checked the cumulative case␣
↪and death count
#for Arizona on the [CDC's Covid Tracker](https://www.cdc.gov/
↪covid-data-tracker/#cases),
#the data suggested that we should aggregate the values for both variables.
#which(df_covid2$state == "Arizona" | df_covid2$state == "arizona")

which(df_covid2$state == "Arizona" | df_covid2$state == "arizona")
df_covid2[3:4,]

az_case_sum = sum(df_covid2$total_case[3:4])
az_death_sum = sum(df_covid2$total_death[3:4])

df_covid2$total_case[3] = az_case_sum
df_covid2$total_death[3] = az_death_sum
df_covid2 = df_covid2[!(df_covid2$state %in% "arizona"), ]
```

2.4 Extreme Outlier Identification

Because each state's policy selection, timeliness of response, and the duration of each policy varies, the range of variables is quite wide.

2.4.1 Covid-19 Metrics

In this section, we will examine **CaseRate_100k, CasesInLast7Days_100k, totalTestResults_100k, and infection_rate** to identify extreme values and determine the most appropriate treatment option.

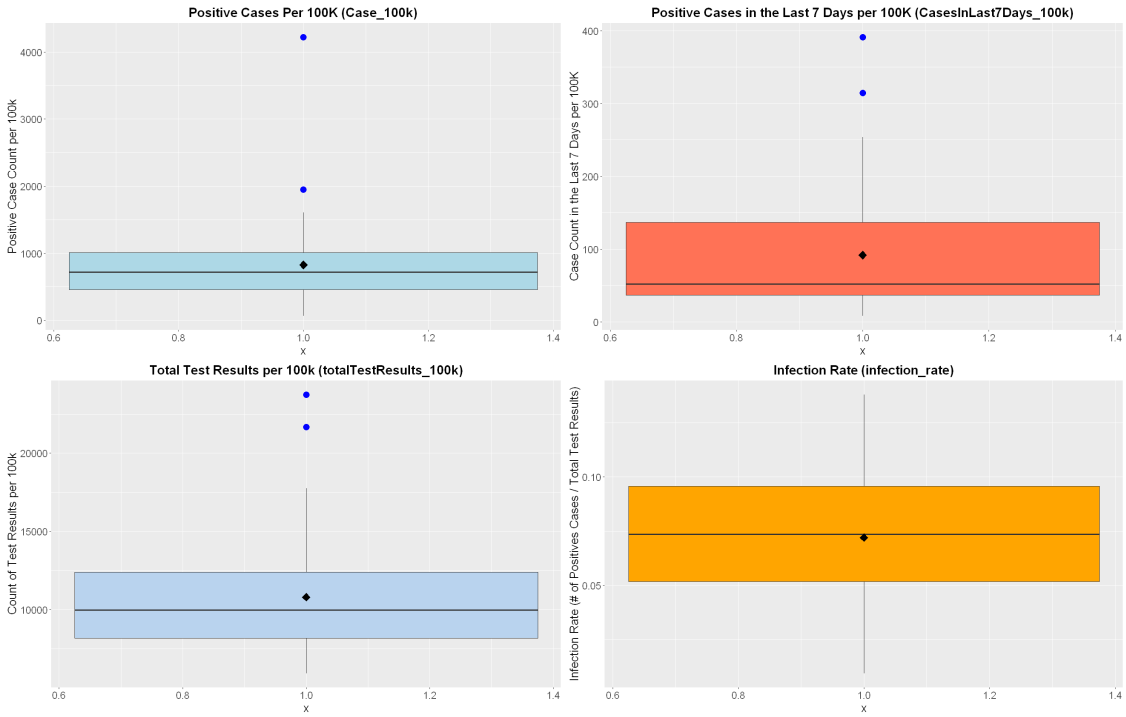
[9]:
```
summary(df_covid2[, c("CaseRate_100k", "CasesInLast7Days_100k",␣
↪"totalTestResults_100k", "infection_rate")])
```

```
 CaseRate_100k    CasesInLast7Days_100k totalTestResults_100k infection_rate
```

```
Min.   :  66.1   Min.   :  8.10        Min.   : 5936         Min.   :0.00930
1st Qu.: 458.6   1st Qu.: 36.65        1st Qu.: 8196         1st Qu.:0.05165
Median : 717.4   Median : 51.80        Median : 9985         Median :0.07360
Mean   : 826.9   Mean   : 91.56        Mean   :10803         Mean   :0.07204
3rd Qu.:1013.2   3rd Qu.:136.70        3rd Qu.:12413         3rd Qu.:0.09570
Max.   :4220.4   Max.   :391.00        Max.   :23735         Max.   :0.13800
```

[10]:
```
options(repr.plot.width = 22, repr.plot.height = 14)
plot_theme =  theme(plot.title = element_text(size = 18, hjust = .5, face = "bold"), axis.text = element_text(size = 14), axis.title = element_text(size = 16),
                    legend.title = element_text(size = 16), legend.text=element_text(size = 14), legend.position = "top",
                    strip.text.y = element_text(size = 16, color = "black", angle = 270))

case100k_box = ggplot(df_covid2, aes(y = CaseRate_100k, x = 1)) +
    geom_boxplot(outlier.color = "blue", outlier.size = 4, notch = FALSE, fill = "lightblue") +
    labs(title = "Positive Cases Per 100K (Case_100k)", y = "Positive Case Count per 100k") +
    stat_summary(fun = mean, geom = "point", shape = 23, size = 4, fill = "black") +
    plot_theme

caselast7days100k_box = ggplot(df_covid2, aes(y = CasesInLast7Days_100k, x = 1)) +
    geom_boxplot(outlier.color = "blue", outlier.size = 4, notch = FALSE, fill = "coral1") +
    labs(title = "Positive Cases in the Last 7 Days per 100K (CasesInLast7Days_100k)", y = "Case Count in the Last 7 Days per 100K") +
    stat_summary(fun = mean, geom = "point", shape = 23, size = 4, fill = "black") +
    plot_theme

testresults100k_box = ggplot(df_covid2, aes(y = totalTestResults_100k, x = 1)) +
    geom_boxplot(outlier.color = "blue", outlier.size = 4, notch = FALSE, fill = "slategray2") +
    labs(title = "Total Test Results per 100k (totalTestResults_100k)", y = "Count of Test Results per 100k") +
    stat_summary(fun = mean, geom = "point", shape = 23, size = 4, fill = "black") +
    plot_theme

infectrate_box = ggplot(df_covid2, aes(y = infection_rate, x = 1)) +
```

```
    geom_boxplot(outlier.color = "blue", outlier.size = 4, notch = FALSE, fill␣
 ↪= "orange") +
    labs(title = "Infection Rate (infection_rate)", y = "Infection Rate (# of␣
 ↪Positives Cases / Total Test Results)") +
    stat_summary(fun = mean, geom = "point", shape = 23, size = 4, fill =␣
 ↪"black") +
    plot_theme

plot_grid(case100k_box, caselast7days100k_box, testresults100k_box,␣
 ↪infectrate_box, ncol = 2, align = "h", rel_widths = c(2, 2, 2))
```



[11]:
```
"Total Test Results per 100k (totalTestResults_100k) outliers in descending␣
 ↪order"
top_n(df_covid2, 2, totalTestResults_100k) %>%␣
 ↪arrange(desc(totalTestResults_100k)) %>% select(state, totalTestResults_100k)
"Positive Case Per 100K (CaseRate_100k) outliers in descending order"
top_n(df_covid2, 2, CaseRate_100k) %>% arrange(desc(CaseRate_100k)) %>%␣
 ↪select(state, CaseRate_100k)
"Positive Cases in the Last 7 Days per 100K (CasesInLast7Days_100k) outliers in␣
 ↪descending order"
top_n(df_covid2, 2, CasesInLast7Days_100k) %>%␣
 ↪arrange(desc(CasesInLast7Days_100k)) %>% select(state, CasesInLast7Days_100k)
```

'Total Test Results per 100k (totalTestResults_100k) outliers in descending order'

A data.frame: 2 × 2

state	totalTestResults_100k
<fct>	<dbl>
Rhode Island	23735.0
New York	21664.9

'Positive Case Per 100K (CaseRate_100k) outliers in descending order'

A data.frame: 2 × 2

state	CaseRate_100k
<fct>	<dbl>
New York	4220.4
New Jersey	1946.5

'Positive Cases in the Last 7 Days per 100K (CasesInLast7Days_100k) outliers in descending order'

A data.frame: 2 × 2

state	CasesInLast7Days_100k
<fct>	<dbl>
Arizona	391.0
Florida	314.5

There are 2 outliers for total test results per 100k, positive cases per 100k, and positive cases in the last 7 days per 100k. Although both New York and Rhode Island have been aggressive in testing, only New York is considered an extreme outlier with respect to positive case count per 100k. Even though these states' Covid-19 related statistics are significantly different than the rest of the country, they still represent real events and we should not dispel them. Thus, we opt to keep them in the dataset.

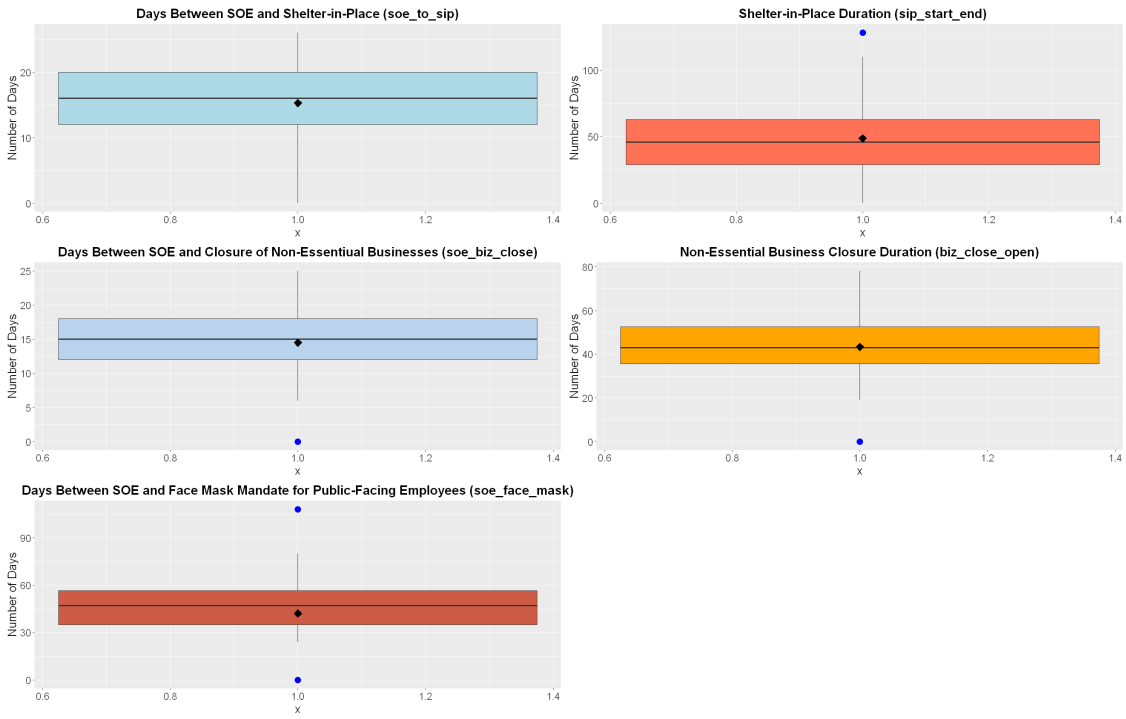
2.4.2 State Response and Policy Duration

In this section, we will examine our main independent variables of interest identify extreme values and determine the most appropriate treatment option: **soe_to_sip, sip_start_end, soe_biz_close, biz_close_open, and soe_face_mask.** Specifically, we hypothesize:

- soe_to_sip: While declaring a state of emergency is the first act to recognize the seriousness of the pandemic, the declaration itself have minimal impact to people's behavior. Measuring the number of days before state legislators followed-up with actionable response such as shelter-in-place can potentially curb infection rate
- sip_start_end: Shelter-in-place reduces contact rate between individuals. Thus, a logical extrapolation is that the longer the duration of shelter-in-place, the less severe infection rate will be
- soe_biz_close: Although governments can issue shelter-in-place, most orders were not enforced. Thus, the cost of defying orders from a legal perspective was low. The closure of all non-essential businesses was another avenue to disincentivize residents from leaving their home.
- biz_close_open: Our belief is that the duration of business closure can impact infection rate; the sooner the closure is lifted, the higher the infection rate because it offers options for people to congregate outside of their homes
- soe_face_mask: Because essential employees are still went to and came in contact with non-familial individuals on a daily basis, their risk level is relatively high. For that reason, we believe the time it took state policymakers to mandate face mask for public-facing employees could impact infection rate

[12]:
```
soetosip_box = ggplot(df_covid2, aes(y = soe_to_sip, x = 1)) +
    geom_boxplot(outlier.color = "blue", outlier.size = 4, notch = FALSE, fill␣
 ↪= "lightblue") +
    labs(title = "Days Between SOE and Shelter-in-Place (soe_to_sip)", y =␣
 ↪"Number of Days") +
    stat_summary(fun = mean, geom = "point", shape = 23, size = 4, fill =␣
 ↪"black") +
    plot_theme

sipstartend_box = ggplot(df_covid2, aes(y = sip_start_end, x = 1)) +
    geom_boxplot(outlier.color = "blue", outlier.size = 4, notch = FALSE, fill␣
 ↪= "coral1") +
    labs(title = "Shelter-in-Place Duration (sip_start_end)", y = "Number of␣
 ↪Days") +
    stat_summary(fun = mean, geom = "point", shape = 23, size = 4, fill =␣
 ↪"black") +
    plot_theme

soebizclose_box = ggplot(df_covid2, aes(y = soe_biz_close, x = 1)) +
    geom_boxplot(outlier.color = "blue", outlier.size = 4, notch = FALSE, fill␣
 ↪= "slategray2") +
    labs(title = "Days Between SOE and Closure of Non-Essentiual Businesses␣
 ↪(soe_biz_close)", y = "Number of Days") +
    stat_summary(fun = mean, geom = "point", shape = 23, size = 4, fill =␣
 ↪"black") +
    plot_theme

bizcloseopen_box = ggplot(df_covid2, aes(y = biz_close_open, x = 1)) +
    geom_boxplot(outlier.color = "blue", outlier.size = 4, notch = FALSE, fill␣
 ↪= "orange") +
    labs(title = "Non-Essential Business Closure Duration (biz_close_open)", y␣
 ↪= "Number of Days") +
    stat_summary(fun = mean, geom = "point", shape = 23, size = 4, fill =␣
 ↪"black") +
    plot_theme

soefacemask_box = ggplot(df_covid2, aes(y = soe_face_mask, x = 1)) +
    geom_boxplot(outlier.color = "blue", outlier.size = 4, notch = FALSE, fill␣
 ↪= "coral3") +
    labs(title = "Days Between SOE and Face Mask Mandate for Public-Facing␣
 ↪Employees (soe_face_mask)", y = "Number of Days") +
    stat_summary(fun = mean, geom = "point", shape = 23, size = 4, fill =␣
 ↪"black") +
    plot_theme
```

```
plot_grid(soetosip_box, sipstartend_box, soebizclose_box, bizcloseopen_box,␣
 ↪soefacemask_box, nrow = 3, align = "h", rel_widths = c(2, 2, 2))
```



[13]:
```
"Shelter-in-Place Length outliers"
top_n(df_covid2, 1, sip_start_end) %>% arrange(desc(sip_start_end)) %>%␣
 ↪select(state, sip_start_end)
"Days Between SOE and Closure of Non-Essentiual Businesses outliers"
top_n(df_covid2, -1, soe_biz_close) %>% arrange(desc(soe_biz_close)) %>%␣
 ↪select(state, soe_biz_close)
"Non-Essential Business Closure Duration outliers"
top_n(df_covid2, -1, biz_close_open) %>% arrange(desc(biz_close_open)) %>%␣
 ↪select(state, biz_close_open)
"Days Between SOE and Face Mask Mandate for Public-Facing Employees outliers"
top_n(df_covid2, 1, soe_face_mask) %>% arrange(desc(soe_face_mask)) %>%␣
 ↪select(state, soe_face_mask)
top_n(df_covid2, -1, soe_face_mask) %>% arrange(desc(soe_face_mask)) %>%␣
 ↪select(state, soe_face_mask)
```

'Shelter-in-Place Length outliers'

A data.frame: 1 × 2

state	sip_start_end
<fct>	<int>
Hawaii	128

'Days Between SOE and Closure of Non-Essentiual Businesses outliers'

A data.frame: 1 × 2

state	soe_biz_close
<fct>	<int>
South Dakota	0

'Non-Essential Business Closure Duration outliers'

A data.frame: 1 × 2

state	biz_close_open
<fct>	<int>
South Dakota	0

'Days Between SOE and Face Mask Mandate for Public-Facing Employees outliers'

A data.frame: 1 × 2

state	soe_face_mask
<fct>	<int>
North Carolina	108

A data.frame: 10 × 2

state	soe_face_mask
<fct>	<int>
Idaho	0
Iowa	0
Kansas	0
Missouri	0
Montana	0
Oklahoma	0
South Carolina	0
South Dakota	0
Tennessee	0
Wisconsin	0

- For shelter-in-place, Hawaii is an outlier. Upon closer examination, its government extended its shelter-in-place order through the end of July (source: Ninth Supplementary Proclamation Related to the Covid-19 Emergency). We're making an assumption here that the order will expire at the end of July 31st and it's still in effect at the time of this analysis (has not been rescinded pre-maturely). For that reason, we will keep Hawaii in our dataset
- For both closure of non-essential businesses and closure duration, South Dakota is an outlier from a statistical perspective. However, the reason is quite simple: South Dakota did not declare such policy and kept all of their businesses open. We believe this represents deviations from the majority of the dataset from both a policy and statistical perspective
- For the number of days between the declaration of state of emergency and the mandate of face coverings for public-facing employees, there are 2 extremes: North Carolina declared a SOE on 3/10, but did not mandate such policy until 6/26. On the contrary, there were 10 states that did not mandate such policy, so the value is 0. While these extremes may impact the goodness of fit of our model, we still believe it's important to keep them in because they represent real legislative choices these states made in their fight against the pandemic.

2.4.3 State Demographics

In this section, we will examine several demographic variables of interest that could impact Covid-19 infection rate: **wunemp_insure_max, pop_dens, perc_under_fed_povline18, perc_atrisk_covid, and all_death_rate18** Specifically, we hypothesize:

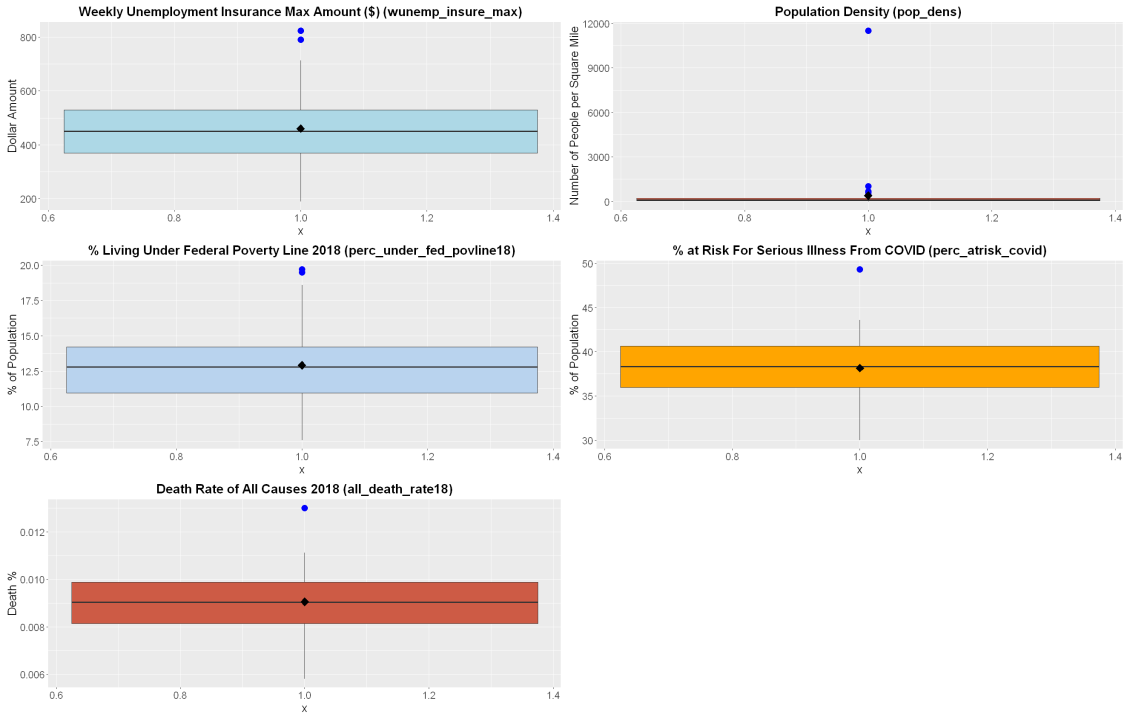
- wunemp_insure_max: We hypothesize a low weekly unemployment insurance amount could

lead to unemployed individuals to seek employment that ultimately increases the risk of exposure and infection

- pop_dens: We believe the more densely populated a state is (i.e. New York), the higher the probability of spread as there are more opportunities for the disease to find a new host
- perc_under_fed_povline18: This variable could correlate with wunemp_insure_max described above, and for the same reason, we believe that poor communities are at more risk of exposure and infection relative to more affluent states because pandemic or no pandemic, these individuals have to work to make ends meet
- perc_atrisk_covid: Although there are reports from health officials that elderly individuals are more prone to serious health complications from Covid-19, our interest lies with infection rate. Based on the latest medical literature, Covid-19 does not discriminate against age
- all_death_rate18: Fatality rate for all causes in 2018 (pre-Covid). Our interest is to use this variable as a crude proxy to the general health of each state's population and the quality of medical care

[14]:
```
insuremax_box = ggplot(df_covid2, aes(y = wunemp_insure_max, x = 1)) +
    geom_boxplot(outlier.color = "blue", outlier.size = 4, notch = FALSE, fill␣
↪= "lightblue") +
    labs(title = "Weekly Unemployment Insurance Max Amount ($)␣
↪(wunemp_insure_max)", y = "Dollar Amount") +
    stat_summary(fun = mean, geom = "point", shape = 23, size = 4, fill =␣
↪"black") +
    plot_theme

popdens_box = ggplot(df_covid2, aes(y = pop_dens, x = 1)) +
    geom_boxplot(outlier.color = "blue", outlier.size = 4, notch = FALSE, fill␣
↪= "coral1") +
    labs(title = "Population Density (pop_dens)", y = "Number of People per␣
↪Square Mile") +
    stat_summary(fun = mean, geom = "point", shape = 23, size = 4, fill =␣
↪"black") +
    plot_theme

underpoverty_box = ggplot(df_covid2, aes(y = perc_under_fed_povline18, x = 1)) +
    geom_boxplot(outlier.color = "blue", outlier.size = 4, notch = FALSE, fill␣
↪= "slategray2") +
    labs(title = "% Living Under Federal Poverty Line 2018␣
↪(perc_under_fed_povline18)", y = "% of Population") +
    stat_summary(fun = mean, geom = "point", shape = 23, size = 4, fill =␣
↪"black") +
    plot_theme

atrisk_box = ggplot(df_covid2, aes(y = perc_atrisk_covid, x = 1)) +
    geom_boxplot(outlier.color = "blue", outlier.size = 4, notch = FALSE, fill␣
↪= "orange") +
    labs(title = "% at Risk For Serious Illness From COVID␣
↪(perc_atrisk_covid)", y = "% of Population") +
```

```
    stat_summary(fun = mean, geom = "point", shape = 23, size = 4, fill =␣
 ↪"black") +
    plot_theme

alldeath18_box = ggplot(df_covid2, aes(y = all_death_rate18, x = 1)) +
    geom_boxplot(outlier.color = "blue", outlier.size = 4, notch = FALSE, fill␣
 ↪= "coral3") +
    labs(title = "Death Rate of All Causes 2018 (all_death_rate18)", y = "Death␣
 ↪%") +
    stat_summary(fun = mean, geom = "point", shape = 23, size = 4, fill =␣
 ↪"black") +
    plot_theme

plot_grid(insuremax_box, popdens_box, underpoverty_box, atrisk_box,␣
 ↪alldeath18_box, nrow = 3, align = "h", rel_widths = c(2, 2, 2))
```



[15]:
```
"Weekly Unemployment Insurance Max Amount outliers"
top_n(df_covid2, 2, wunemp_insure_max) %>% arrange(desc(wunemp_insure_max)) %>%␣
 ↪select(1:3)
"Population Density outliers"
top_n(df_covid2, 2, pop_dens) %>% arrange(desc(pop_dens)) %>% select(state,␣
 ↪pop_dens)
"% Living Under Federal Poverty Line 2018 outliers"
```

```
top_n(df_covid2, 2, perc_under_fed_povline18) %>%␣
 ↪arrange(desc(perc_under_fed_povline18)) %>% select(state,␣
 ↪perc_under_fed_povline18)
"% at Risk For Serious Illness From COVID (perc_atrisk_covid) outliers"
top_n(df_covid2, 1, perc_atrisk_covid) %>% arrange(desc(perc_atrisk_covid)) %>%␣
 ↪select(state, perc_atrisk_covid)
"Death Rate All Causes 2018 outliers"
top_n(df_covid2, 1, all_death_rate18) %>% arrange(desc(all_death_rate18)) %>%␣
 ↪select(state, all_death_rate18)
```

'Weekly Unemployment Insurance Max Amount outliers'

A data.frame: 2 × 3

party2016	party2016_votepercent	state
<fct>	<dbl>	<fct>
Democrat	0.5905	Massachusetts
Democrat	0.5254	Washington

'Population Density outliers'

A data.frame: 2 × 2

state	pop_dens
<fct>	<dbl>
District of Columbia	11496.81
New Jersey	1021.27

'% Living Under Federal Poverty Line 2018 outliers'

A data.frame: 2 × 2

state	perc_under_fed_povline18
<fct>	<dbl>
Mississippi	19.7
New Mexico	19.5

'% at Risk For Serious Illness From COVID (perc_atrisk_covid) outliers'

A data.frame: 1 × 2

state	perc_atrisk_covid
<fct>	<dbl>
West Virginia	49.3

'Death Rate All Causes 2018 outliers'

A data.frame: 1 × 2

state	all_death_rate18
<fct>	<dbl>
West Virginia	0.01300121

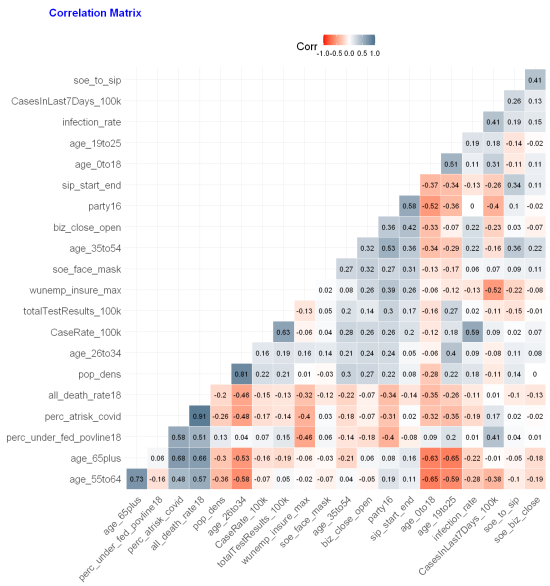
While all of the examined state demographic variables have outliers, the only of which stands out is District of Columbia's population density. More specifically, its population density is more than 11x higher than the next most densely populated state of New Jersey. Washington D.C.'s infection rate is slightly below 10%, but still above the nation's median and average. Similar to the previous set of variables we analyzed, we opt to keep them in the dataset as they represent real-world phenomenas.

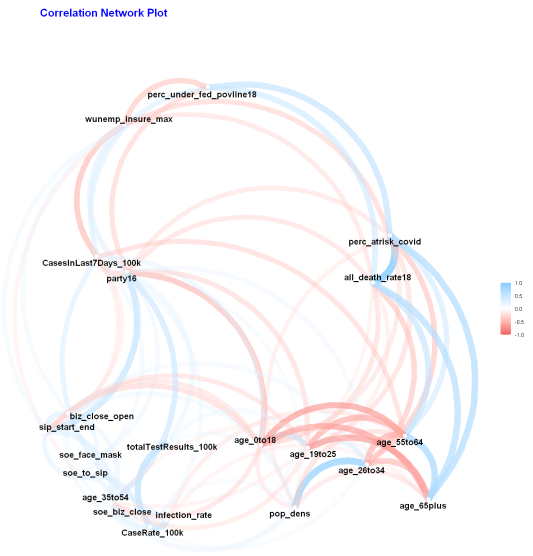
2.5 Correlation Matrix and Plot

To conclude the initial data exploration, we've developed a matrix and network plot to help us quickly identify both the direction and strength of correlation between variables.

[16]:
```
#correlation matrix table
corr_vars = c("infection_rate", "party16", "CasesInLast7Days_100k",␣
 ↪"CaseRate_100k", "totalTestResults_100k","soe_to_sip", "sip_start_end",␣
 ↪"soe_biz_close", "biz_close_open", "soe_face_mask",
            "wunemp_insure_max", "pop_dens", "perc_under_fed_povline18",␣
 ↪"perc_atrisk_covid", "all_death_rate18", "age_0to18", "age_19to25",␣
 ↪"age_26to34", "age_35to54", "age_55to64", "age_65plus")

corr_mat = round(cor(df_covid2[, corr_vars], method = "pearson"),2)

#correlation matrix visual
corr_matrix = ggcorrplot(corr_mat, hc.order = TRUE, type = "lower", outline.col␣
 ↪= "white",  colors = c("red", "white", "skyblue4"), lab = TRUE, tl.cex = 16)␣
 ↪+
theme(legend.title = element_text(size = 18), legend.text = element_text(size =␣
 ↪12), legend.position = "top")

#network correlation plot
corr_data  = data.frame(df_covid2)
corr_data  = corr_data[, corr_vars]
corr_plot = corr_data %>% correlate(method = "pearson") %>%␣
 ↪network_plot(min_cor = .2)

#side-by-side plot
options(repr.plot.width = 26, repr.plot.height = 14)
plot_grid(corr_matrix, NULL, corr_plot, nrow = 1, rel_widths = c(2, .1, 2),␣
 ↪labels = c("Correlation Matrix", "", "Correlation Network Plot"), label_size␣
 ↪= 18, label_colour = "blue")
```

```
Correlation method: 'pearson'
Missing treated using: 'pairwise.complete.obs'
```





From the above matrix, we identify the following noteworthy relationships:

1. **biz_close_open** has the strongest positive correlation with infection rate from a policy perspective. The matrix indicates that the longer non-essential businesses shutdown, the higher the infection rate will be. This goes against our initial intuition because one would think that shutting down non-essential businesses would disincentivize people from going out beyond essential needs. However, this does not seem to be the case
2. **sip_start_end** has a negative correlation with infection rate, which means the longer the duration, the lower the infection rate. This makes sense, since the main purpose of shelter-in-place is to reduce face-to-face contact of individuals
3. **soe_to_sip** has a positive correlation to infection rate, which means that the longer a state waited to enact shelter-in-place after a state of emergency was declared, the higher the infection rate
4. **soe_biz_close** has a positive correlation to infection rate, which means the longer a state waited to close non-essential businesses after the state of emergency was declared, the higher the infection rate
5. **age_35to54** has a positive correlation with infection rate, which means that higher proportion of population that falls into this age interval is likely to increase infection rate. This could be due to both lifestyle/behavioral and health related dimensions
6. **perc_atrisk_covid** has a negative correlation with infection rate, which is counter-intuitive. We would think that the higher the proportion of at-risk individuals for serious illness due to Covid-19, the higher the infection rate, which does not seem to be the case. One explanation is that states with higher percentage of at risk population (i.e. elderly) have taken specific measures to ensure those groups are protected as much as possible (i.e. no family visits to retirement homes)
7. **pop_dens** has the second highest positive correlation outside of the Covid-19 related measures. Our intuition tells us that the more densely populated a state is, the higher the probability of infection due to proximity. Additionally, infection rate could increase at an

accelerated pace due to the large volume of new hosts within a confined space.

8. **age_55to64 and age_65plus** has negative correlation with infection rate. While seniors are more at risk from a fatality perspective, the matrix above indicate that they are less likely to contract the disease. One explanation is that because the fatality rate is relatively high among the elderly, both states and individuals have taken extra pre-caution to limit contact and exposure, thus, reducing the likelihood of contraction.

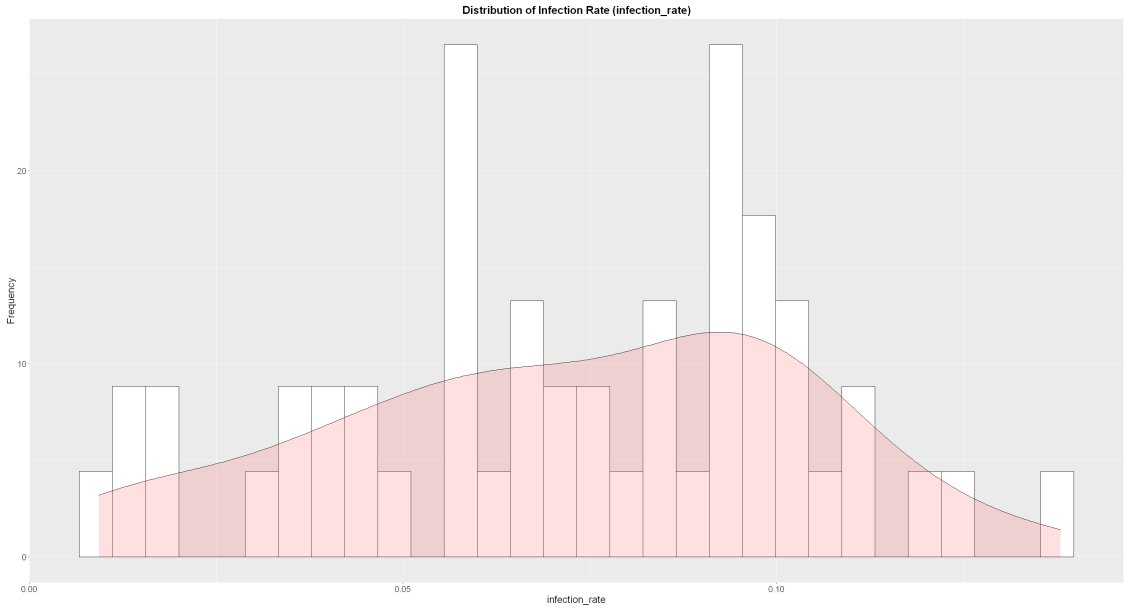
3 Modeling Process

3.1 Dependent Variable

The dependent variable of our analysis is infection_rate, which is defined as:

$$InfectionRate = \frac{totalCases}{totalTestResults}$$

[17]:
```
#original infection_rate
infectrate_org = ggplot(df_covid2, aes(infection_rate)) +
    geom_histogram(aes(y =..density..), bins = 30, color = "black", fill =␣
 ↪"white") +
    geom_density(alpha =.2, fill = "#FF6666") +
    labs(title = "Distribution of Infection Rate (infection_rate)", y =␣
 ↪"Frequency") +
    plot_theme

infectrate_org
```



Here we inspect the distribution of infection_rate for normality by comparing the original distribution against its log-transformed version (see plots below). Upon further analysis, we decided against the log transformation as the original distribution had more semblance to a normal distribution than the log-transformed version (shows symptoms of left-skew).

[7]:
```
shapiro.test(df_covid2$infection_rate)
```

```
Shapiro-Wilk normality test

data:  df_covid2$infection_rate
W = 0.96623, p-value = 0.1457
```

Here we use the Shapiro-Wilk normality test to assess if the distribution is 0. Specifically, if the p-value > 0.05, then it implies that the distribution of the data are not significantly different from the normal distribution. Because the p-value is above 0.05, we cannot reject the null hypothesis that the distribution does not resemble normality

3.2 Model 1

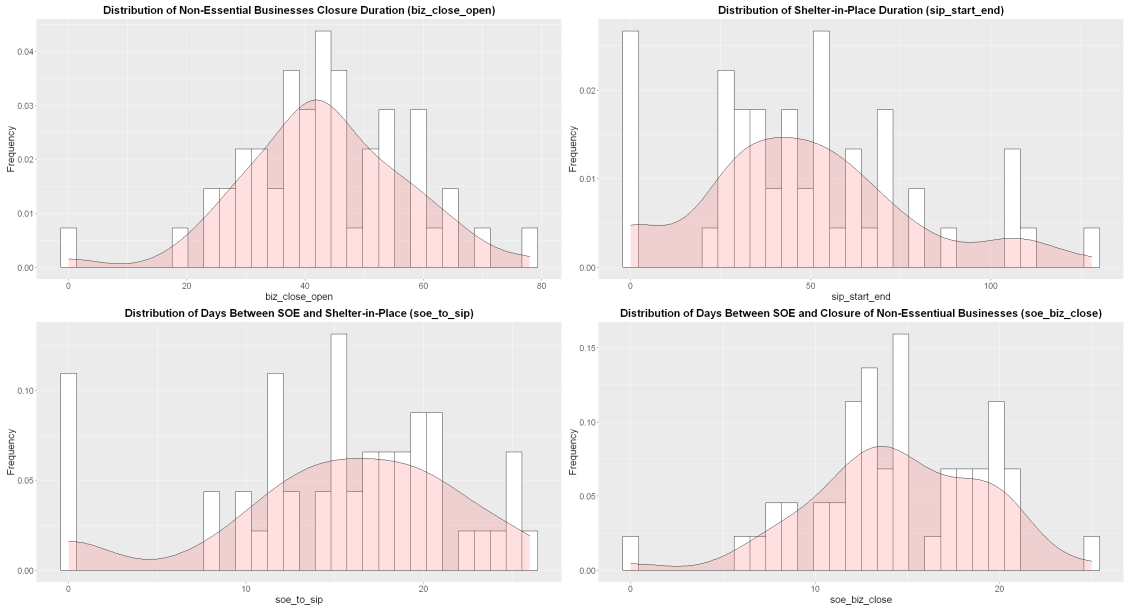
3.2.1 Independent Variable Selection

Our first model will focus on the timeliness and duration of counter-measures/policies put in place to combat the spread of Covid-19, which there are 4 such variables of interest for the baseline model: - **biz_close_open:** Duration of non-essential businesses closure - **sip_start_end:** Duration of shelter-in-place order - **soe_to_sip:** Number of days shelter-in-place was ordered after the state of emergency was declared - **soe_biz_close:** Number of days closure of non-essential businesses was ordered after the state of emergency was declared

Next, we plot the distribution of each independent variable to assess for normality.

[19]:
```
bizcloseopen_org = ggplot(df_covid2, aes(biz_close_open)) +
    geom_histogram(aes(y =..density..), bins = 30, color = "black", fill =␣
 ↪"white") +
    geom_density(alpha =.2, fill = "#FF6666") +
    labs(title = "Distribution of Non-Essential Businesses Closure Duration␣
 ↪(biz_close_open)", y = "Frequency") +
    plot_theme

sipstartend_org = ggplot(df_covid2, aes(sip_start_end)) +
    geom_histogram(aes(y =..density..), bins = 30, color = "black", fill =␣
 ↪"white") +
    geom_density(alpha =.2, fill = "#FF6666") +
    labs(title = "Distribution of Shelter-in-Place Duration (sip_start_end)", y␣
 ↪= "Frequency") +
    plot_theme

soetosip_org = ggplot(df_covid2, aes(soe_to_sip)) +
```

```
    geom_histogram(aes(y =..density..), bins = 30, color = "black", fill =␣
 ↪"white") +
    geom_density(alpha =.2, fill = "#FF6666") +
    labs(title = "Distribution of Days Between SOE and Shelter-in-Place␣
 ↪(soe_to_sip)", y = "Frequency") +
    plot_theme

soebizclose_org = ggplot(df_covid2, aes(soe_biz_close)) +
    geom_histogram(aes(y =..density..), bins = 30, color = "black", fill =␣
 ↪"white") +
    geom_density(alpha =.2, fill = "#FF6666") +
    labs(title = "Distribution of Days Between SOE and Closure of␣
 ↪Non-Essentiual Businesses (soe_biz_close)", y = "Frequency") +
    plot_theme

plot_grid(bizcloseopen_org, sipstartend_org, soetosip_org, soebizclose_org,␣
 ↪nrow = 2, align = "v", rel_widths = c(2, 2, 2))
```



[20]:
```
shapiro.test(df_covid2$biz_close_open)
shapiro.test(df_covid2$sip_start_end)
shapiro.test(df_covid2$soe_to_sip)
shapiro.test(df_covid2$soe_biz_close)
```

```
Shapiro-Wilk normality test

data:  df_covid2$biz_close_open
W = 0.98442, p-value = 0.7357
```

```
Shapiro-Wilk normality test

data:  df_covid2$sip_start_end
W = 0.95193, p-value = 0.03797


Shapiro-Wilk normality test

data:  df_covid2$soe_to_sip
W = 0.92188, p-value = 0.002463


Shapiro-Wilk normality test

data:  df_covid2$soe_biz_close
W = 0.97522, p-value = 0.3595
```

Based on both the histograms and Shapiro-Wilk test results, sip_start_end and soe_to_sip exhibit signs of non-normality. This is partially due to 4 states that did not shelter-in-place orders. Because the dataset contains more than 30 samples, we can invoke the Central Limit Theorem (CLT) and proceed.

3.2.2 Model Execution and Interpretation

The model output below does not seem to suggest that the timeliness of non-essential business closure (soe_biz_close) was a significant factor with respect to infection rate. We will discard this variable and assess for impact on the remaining variables.

[21]:
```
model1a = lm(infection_rate ~ biz_close_open + sip_start_end + soe_to_sip +␣
 ↪soe_biz_close, data = df_covid2)
summary(model1a)
```

```
Call:
lm(formula = infection_rate ~ biz_close_open + sip_start_end +
    soe_to_sip + soe_biz_close, data = df_covid2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.061549 -0.018384 -0.000136  0.021102  0.061934

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.0245887  0.0197388   1.246   0.2192
```

```
biz_close_open  0.0008624  0.0003313   2.603   0.0124 *
sip_start_end  -0.0004083  0.0001620  -2.520   0.0153 *
soe_to_sip      0.0012277  0.0007129   1.722   0.0918 .
soe_biz_close   0.0007724  0.0009734   0.794   0.4315
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02935 on 46 degrees of freedom
Multiple R-squared:  0.2023,Adjusted R-squared:  0.1329
F-statistic: 2.916 on 4 and 46 DF,  p-value: 0.03124
```

After removing soe_biz_close, the remaining 3 independent variables became statistically significant. However, public policies and their durations was only able to explain 15% of the variations in infection rate per the R^2. It's likely that we are missing other independent variables, either within the dataset or omitted.

Interpretation: The model output suggest that for every day that non-essential businesses are closed, infection rate increases by 0.00084 percentage points holding all other variables constant. For the duration of shelter-in-place order, the model suggest that for each incremental day that the order is in effect, infection rate is expected to decrease by 0.0004 percentage points holding all other variables constant. Moreover, for each day that state waited to enact shelter-in-place after a state of emergency was declared, the infection rate is expected to increase by 0.0014 percentage points holding all other variables constant. Although the percentage points seem small, we must remember that these figures are applied to millions of people, which is not insignificant!

[22]:
```
model1 = lm(infection_rate ~ biz_close_open + sip_start_end + soe_to_sip, data␣
 ↪= df_covid2)
summary(model1)
```

```
Call:
lm(formula = infection_rate ~ biz_close_open + sip_start_end +
    soe_to_sip, data = df_covid2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.060360 -0.017202 -0.001689  0.020600  0.061664

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.0333897  0.0162637   2.053   0.0457 *
biz_close_open  0.0008391  0.0003287   2.553   0.0140 *
sip_start_end  -0.0004072  0.0001614  -2.523   0.0151 *
soe_to_sip      0.0014476  0.0006542   2.213   0.0318 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02923 on 47 degrees of freedom
```

```
Multiple R-squared:  0.1914,Adjusted R-squared:  0.1398
F-statistic: 3.708 on 3 and 47 DF,  p-value: 0.01784
```

Only the coefficients of sip_start_end and soe_sip are statistically significant when we examine the in the output below while the p-value for biz_close_open is slightly above 0.05.

[23]:
```
coeftest(model1, vcov = vcovHC, level = 0.05)
```

```
t test of coefficients:

                 Estimate  Std. Error t value Pr(>|t|)
(Intercept)     0.03338975  0.02163764  1.5431  0.12951
biz_close_open  0.00083911  0.00043324  1.9368  0.05879 .
sip_start_end  -0.00040720  0.00017586 -2.3155  0.02500 *
soe_to_sip      0.00144759  0.00066684  2.1708  0.03503 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3.2.3 Model Diagnostics

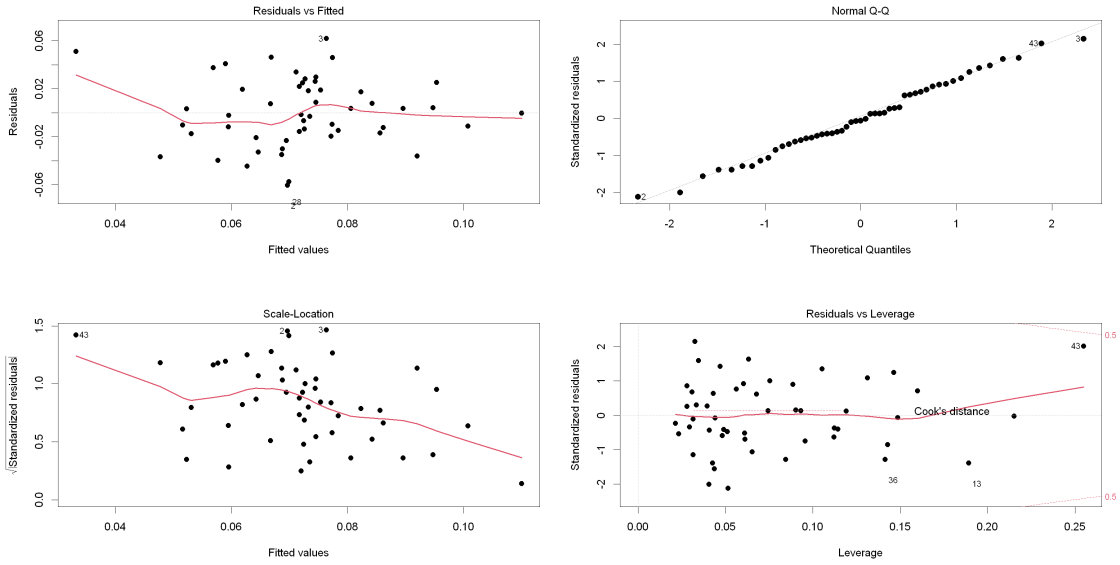
We use a combination of visual inspections and formal statistical tests to check on the 6 Classical Linear Model (CLM) assumptions:

- **MLR.1 Linearity in Parameters:** For model 1, we assume linearity in parameters by default and MLR.1 holds. We have not transformed our dependent variable infection_rate or any of the independent variables.
- **MLR.2 Random Sampling:** We unfortunately cannot assume that MLR.2 holds as they may not be true random samples. Individuals who are sick are more likely to get tested while Individuals who may be sick, but are asymptomatic, may be afraid of getting testing due to social stigma. However, the p-value of 0.338 from the Durbin-Watson below indicates that there is no autocorrelation in the dataset. Thus, we can still perform linear modeling because we assume the sample is still representative of the underlying population.
- **MLR.3 No Perfect Collinearity:** We used the Variance Inflation Score (VIF) to determine if there are symptoms of multicollinearity among our independent variables (see VIF score below), which is not easily identifiable through the diagnostic plots.
  - **1. Residuals vs. Fitted:** The plot reveals a slight non-linear pattern in the plot below. However, because the bend is slight in nature, we do not believe that it violates the zero conditional mean of errors assumption. Thus, we believe **MLR.4** holds
  - **2. Normal Q-Q:** This plot shows if residuals are normally distributed if it follows a relatively straight line. From the plot below, the bulk of the error terms seem to follow the straight line, which suggests a fairly normal distribution. Additionally, the Shapiro-Wilk normality test returned a p-value of 0.9295 (see test below plots), further supporting the evidence that the normality of errors/residuals assumption holds (**MLR.6**)
  - **3. Scale-Location:** This plot shows if residuals are spread equally along the ranges of explanatory variables. From the plot below, the red line is downward sloping but shows

signs of heteroskedasticity as the spread of errors are largely concentrated in the center. We then use the Breusch-Pagan test (see test below plots) to help us determine if **MLR.5** holds. The p-value is 0.0896, slightly bigger than the 0.05 threshold. Thus, while we can say that homoscedasticity holds, it does show symptoms of heteroskedasticity, which means we should not use the standard errors provided in the regression output, but the heteroskedastic-robust errors computed in the previous section.

- **4. Residuals vs. Leverage:** This plot helps us find influential or high leverage points if any. An influential value is a value can alter the results of the regression analysis and is associated with a large residual. In our specific plot below, all points are within the bounds of Cook's distance lines although observation 43 is nearing the edge.

[24]:
```
par(mfrow = c(2, 2), cex = 1.5)
plot(model1, cex.caption = 1.5, cex.axis = 1, cex.lab = 1, pch = 20, cex = 1.5,␣
↪lwd = 3)
```



Shapiro-Wilk normality test (MLR.6): If p-value > 0.05, then it implies that the distribution of the data are not significantly different from the normal distribution.

[25]:
```
shapiro.test(model1$residuals)
```

```
Shapiro-Wilk normality test

data:  model1$residuals
W = 0.9895, p-value = 0.9295
```

Breusch-Pagan test (MLR.5): If p-value < 0.05, then it implies the null hypothesis of homoscedasticity should be rejected.

[26]:
```
bptest(model1)
```

```
studentized Breusch-Pagan test

data:  model1
BP = 6.5016, df = 3, p-value = 0.0896
```

Variance Inflation Factor (VIF) scores (MLR.3): Quantifies the extent of correlation between one predictor and the other predictor variables in a regression model; the higher the value, the greater the correlation of the variable with other variables. Values of more than 5 are sometimes regarded as problematic.

From the output below, all of our predictor variables are well below the 5 threshold, so we can claim that there is no perfect collinearity (**MLR.3**) in model 1.

[ ]:
```
car::vif(model1)
```

Variable	VIF	x<5?
biz_close_open	1.23	TRUE
sip_start_end	1.39	TRUE
soe_to_sip	1.15	TRUE

Durbin-Watson test (MLR.2): The Durbin Watson statistic is a test for autocorrelation in the residuals and we can use the output to help assess our random sampling assumption. Specifically, a p-value greater than .05 means that we cannot reject the null hypothesis that there is no autocorrelation among the residuals.

[58]:
```
durbinWatsonTest(model1)
```

```
 lag Autocorrelation D-W Statistic p-value
   1      -0.1538372      2.286713   0.338
 Alternative hypothesis: rho != 0
```

Based on our review of the CLM assumptions, we believe that model 1 is an unbiased estimator of infection rate. However, the model has a poor fit and does not have much explanatory power. Thus, we will add additional covariates that will improve model fit.

3.3 Model 2

3.3.1 Independent Variable Selection

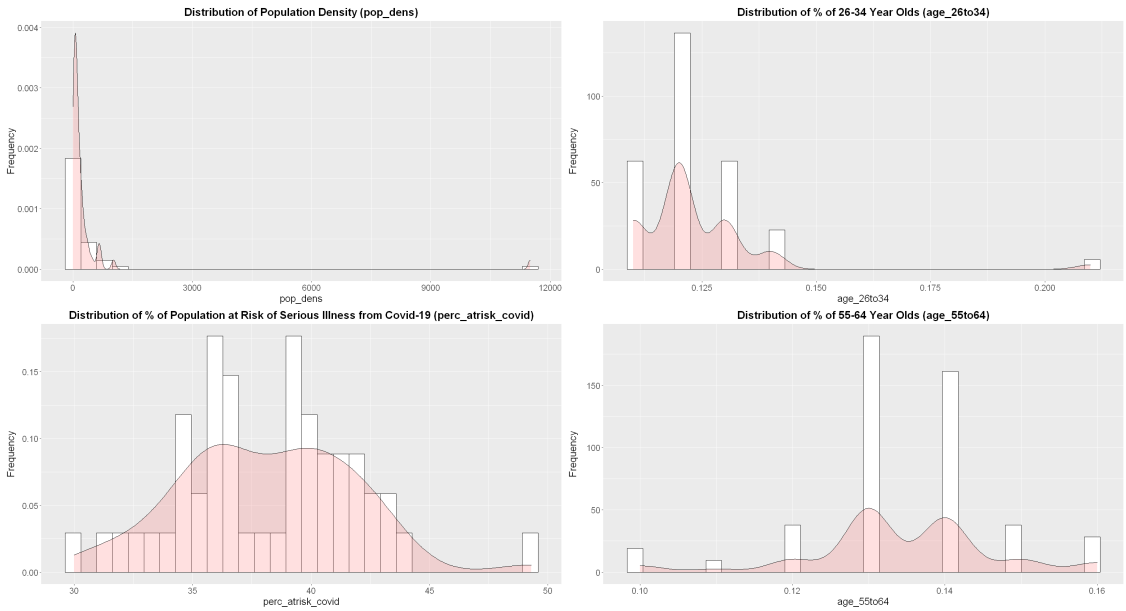
The second iteration of our model builds off of our initial model. In addition to the 3 policy related variables in model 1, we also include state characteristic and demographic variables that we believe can improve model fit and explanatory power. We favor the below variables over potentially other key factors (i.e. % under poverty line and maximum unemployment insurance amount) because they not provided additional boost to explanatory power, but also of practical value.

- **pop_dens:** Our intuition tells us that the more densely populated the state is, the higher the probability of infection due to proximity and human contact
- **age_26to34:** While fatality rate has been highest among the elderly population, emerging data suggest that young adults are more prone to infection, which could be due to lifestyle and behavioral factors
- **age_55to64:** In addition to age_26to34, we also include age_55to64 as this variable has the strongest negative correlation with infection rate. One rationale is that individuals within this age group has taken extra precaution and counter-measures to guard against potential exposure. Because this variable is highly correlated with age_65plus, we have excluded it from the model
- **perc_atrisk_covid:** This variable measure the proportion of the state's population that are risk for serious illness from Covid-19. Unsurprisingly, it's also highly correlated with age_65plus, which we've exlucded from the mode for the same reason as explained in the age_55to64 variable.

Next, we plot the distribution of each new independent variable to assess for normality.

[28]:
```
popdens_org = ggplot(df_covid2, aes(pop_dens)) +
    geom_histogram(aes(y =..density..), bins = 30, color = "black", fill =␣
 ↪"white") +
    geom_density(alpha =.2, fill = "#FF6666") +
    labs(title = "Distribution of Population Density (pop_dens)", y =␣
 ↪"Frequency") +
    plot_theme

age26to34_org = ggplot(df_covid2, aes(age_26to34)) +
    geom_histogram(aes(y =..density..), bins = 30, color = "black", fill =␣
 ↪"white") +
    geom_density(alpha =.2, fill = "#FF6666") +
    labs(title = "Distribution of % of 26-34 Year Olds (age_26to34)", y =␣
 ↪"Frequency") +
    plot_theme

percatriskcovid_org = ggplot(df_covid2, aes(perc_atrisk_covid)) +
    geom_histogram(aes(y =..density..), bins = 30, color = "black", fill =␣
 ↪"white") +
    geom_density(alpha =.2, fill = "#FF6666") +
    labs(title = "Distribution of % of Population at Risk of Serious Illness␣
 ↪from Covid-19 (perc_atrisk_covid)", y = "Frequency") +
    plot_theme

age55to64_org = ggplot(df_covid2, aes(age_55to64)) +
    geom_histogram(aes(y =..density..), bins = 30, color = "black", fill =␣
 ↪"white") +
    geom_density(alpha =.2, fill = "#FF6666") +
    labs(title = "Distribution of % of 55-64 Year Olds (age_55to64)", y =␣
 ↪"Frequency") +
```

```
    plot_theme

plot_grid(popdens_org, age26to34_org, percatriskcovid_org, age55to64_org, nrow␣
 ↪= 2, align = "v", rel_widths = c(2, 2, 2))
```



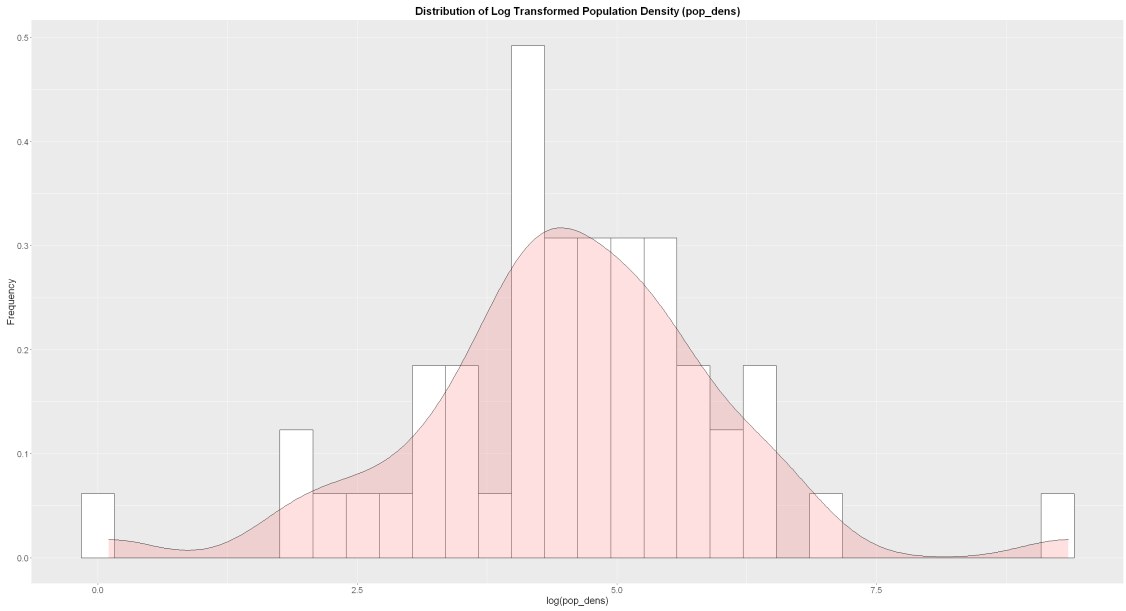
[29]:
```
shapiro.test(df_covid2$pop_dens)
shapiro.test(df_covid2$age_26to34)
shapiro.test(df_covid2$perc_atrisk_covid)
shapiro.test(df_covid2$age_55to64)
```

```
Shapiro-Wilk normality test

data:  df_covid2$pop_dens
W = 0.20026, p-value = 3.682e-15


Shapiro-Wilk normality test

data:  df_covid2$age_26to34
W = 0.60782, p-value = 1.967e-10


Shapiro-Wilk normality test

data:  df_covid2$perc_atrisk_covid
W = 0.98114, p-value = 0.5882
```

```
Shapiro-Wilk normality test

data:  df_covid2$age_55to64
W = 0.89616, p-value = 0.0003139
```

Based on the histograms and Shapiro-Wilk tests, population density should be transformed. After doing so, through both visual inspection coupled with the Shapiro-Wilk test (p-value: 0.1826), population density resembles a normal distribution. We also tried log transforming age_26to34 and age_55to64, but it did not improve normality. However, because we have more than 30 samples in our dataset, we can invoke the CLT to proceed with our model building process.

[30]:
```
popdens_log = ggplot(df_covid2, aes(log(pop_dens))) +
    geom_histogram(aes(y =..density..), bins = 30, color = "black", fill =␣
↪"white") +
    geom_density(alpha =.2, fill = "#FF6666") +
    labs(title = "Distribution of Log Transformed Population Density␣
↪(pop_dens)", y = "Frequency") +
    plot_theme

popdens_log
```



[31]:
```
shapiro.test(log(df_covid2$pop_dens))
```

```
Shapiro-Wilk normality test
```

```
data:  log(df_covid2$pop_dens)
W = 0.96801, p-value = 0.1826
```

3.3.2 Model Execution and Interpretation

By adding the state characteristic and demographic data, we more than doubled our adjusted R^2 to 0.3561 while slightly reducing the standard error. However, several of the original policy related variables are no longer statistically significant. This isn't surprising, since policy duration variables were only able to explain 13% of the variations in inspection rate as observed in model 1. Next, we will remove statistically insignificant variables in a step-wise manner to maximize fit and parsimony.

[32]:
```
model2a = lm(infection_rate ~ biz_close_open + sip_start_end + soe_to_sip +␣
↪soe_biz_close + log(pop_dens) + perc_atrisk_covid + age_55to64, data =␣
↪df_covid2)
summary(model2a)
```

```
Call:
lm(formula = infection_rate ~ biz_close_open + sip_start_end +
    soe_to_sip + soe_biz_close + log(pop_dens) + perc_atrisk_covid +
    age_55to64, data = df_covid2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.052887 -0.014840 -0.000495  0.017370  0.064361

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        1.049e-01  5.082e-02   2.063 0.045139 *
biz_close_open     2.005e-04  3.245e-04   0.618 0.539853
sip_start_end     -4.253e-04  1.432e-04  -2.970 0.004852 **
soe_to_sip         5.161e-04  6.402e-04   0.806 0.424571
soe_biz_close     -1.660e-06  8.645e-04  -0.002 0.998477
log(pop_dens)      1.217e-02  3.038e-03   4.004 0.000242 ***
perc_atrisk_covid -1.020e-03  1.117e-03  -0.913 0.366253
age_55to64        -3.357e-01  3.447e-01  -0.974 0.335611
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02529 on 43 degrees of freedom
Multiple R-squared:  0.4463,Adjusted R-squared:  0.3561
F-statistic: 4.951 on 7 and 43 DF,  p-value: 0.0003625
```

By removing biz_close_open, soe_to_sip, and soe_biz_close, we were able to significantly improve model fit while decreasing the standard error. The resulting model can be seen below (model2b)

Interpretation: The duration of shelter-in-place is statistically significant, and for every incremental day the order is in effect, we can expect an 0.00039 percentage point decrease in infection rate holding all other variables constant (0.1% less than model 1). Additionally, we can interpret the log transformed population density as for every 1% increase in population density, we can expect a 0.017% increase in infection rate.

For every 1 unit increase perc_atrisk_covid, we can expect a 0.22% decrease in infection rate holding all other variables constant. This seems counter-intuitive at first, but one hypothesis is that because individuals in the high risk category have taken extra pre-cautions to mitigate contraction of the disease through more stringent social distancing protocols.

Counter to our initial hypothesis, individuals between 26 to 34 years old exhibit a negative impact to infection rate, which can be interpreted as for every 1 unit increase in age_26to34, we can expect infection rate to decrease by 1.1 percentage point holding all other variables constant. This is particularly interesting because news reports have identified this age group to carry the highest risk of catching Covid-19.

A similar narrative can be told for those in the 55 - 64 age group, which we can expect a 0.97 percentage point decrease in infection rate. This seems to support our initial belief that because the older population group is more susceptible to Covid-19 induced fatality, they may be taking extra steps to protect themselves from exposure.

[33]:
```
model2 = lm(infection_rate ~ sip_start_end + log(pop_dens) + perc_atrisk_covid␣
 ↪+ age_26to34 + age_55to64, data = df_covid2)
summary(model2)
```

```
Call:
lm(formula = infection_rate ~ sip_start_end + log(pop_dens) +
    perc_atrisk_covid + age_26to34 + age_55to64, data = df_covid2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.041395 -0.013395 -0.000098  0.013643  0.056476

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.3616349  0.0749245   4.827 1.63e-05 ***
sip_start_end     -0.0003887  0.0001113  -3.493 0.001085 **
log(pop_dens)      0.0172823  0.0024522   7.048 8.66e-09 ***
perc_atrisk_covid -0.0021855  0.0010043  -2.176 0.034837 *
age_26to34        -1.1003884  0.2884785  -3.814 0.000413 ***
age_55to64        -0.9692330  0.3248753  -2.983 0.004593 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02171 on 45 degrees of freedom
Multiple R-squared:  0.5729,Adjusted R-squared:  0.5255
F-statistic: 12.07 on 5 and 45 DF,  p-value: 1.923e-07
```

When we examine the heteroskedastic-robust errors in the output below, perc_atrisk_covid is no longer statistically significant. If we remove perc_atrisk_covid from the model, then the adjusted R^2 decreases while the standard error increases (see model2b output below)

[34]:
```
coeftest(model2, vcov = vcovHC, level = 0.05)
```

```
t test of coefficients:

                     Estimate  Std. Error t value  Pr(>|t|)
(Intercept)        0.36163492  0.09793833  3.6925 0.0005984 ***
sip_start_end     -0.00038869  0.00013572 -2.8639 0.0063353 **
log(pop_dens)      0.01728228  0.00172137 10.0399  4.59e-13 ***
perc_atrisk_covid -0.00218553  0.00118417 -1.8456 0.0715315 .
age_26to34        -1.10038843  0.27461410 -4.0070 0.0002282 ***
age_55to64        -0.96923299  0.44645824 -2.1709 0.0352496 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

When we remove both perc_atrisk_covid, the adjusted R^2 decreases by roughly .039 as seen in the output below

[35]:
```
model2b = lm(infection_rate ~ sip_start_end + log(pop_dens) + age_26to34 +␣
 ↪age_55to64, data = df_covid2)
summary(model2b)
```

```
Call:
lm(formula = infection_rate ~ sip_start_end + log(pop_dens) +
    age_26to34 + age_55to64, data = df_covid2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.046971 -0.011252  0.001210  0.009754  0.052067

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.2821378  0.0680190   4.148 0.000143 ***
sip_start_end -0.0003730  0.0001155  -3.230 0.002286 **
log(pop_dens)  0.0163482  0.0025105   6.512 4.93e-08 ***
age_26to34    -0.9022703  0.2846329  -3.170 0.002712 **
age_55to64    -1.1540190  0.3260650  -3.539 0.000931 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02257 on 46 degrees of freedom
Multiple R-squared:  0.528,Adjusted R-squared:  0.4869
```

```
F-statistic: 12.86 on 4 and 46 DF,  p-value: 4.168e-07
```

The heteroskedastic-robust errors also confirms the significance of the remaining 4 variables

[36]:
```
coeftest(model2b, vcov = vcovHC, level = 0.05)
```

```
t test of coefficients:

                 Estimate  Std. Error t value  Pr(>|t|)
(Intercept)    0.28213776  0.08179285  3.4494  0.001213 **
sip_start_end -0.00037303  0.00014547 -2.5643  0.013668 *
log(pop_dens)  0.01634816  0.00191131  8.5534 4.582e-11 ***
age_26to34    -0.90227032  0.34008563 -2.6531  0.010912 *
age_55to64    -1.15401895  0.37685848 -3.0622  0.003664 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3.3.3 Model Diagnostics

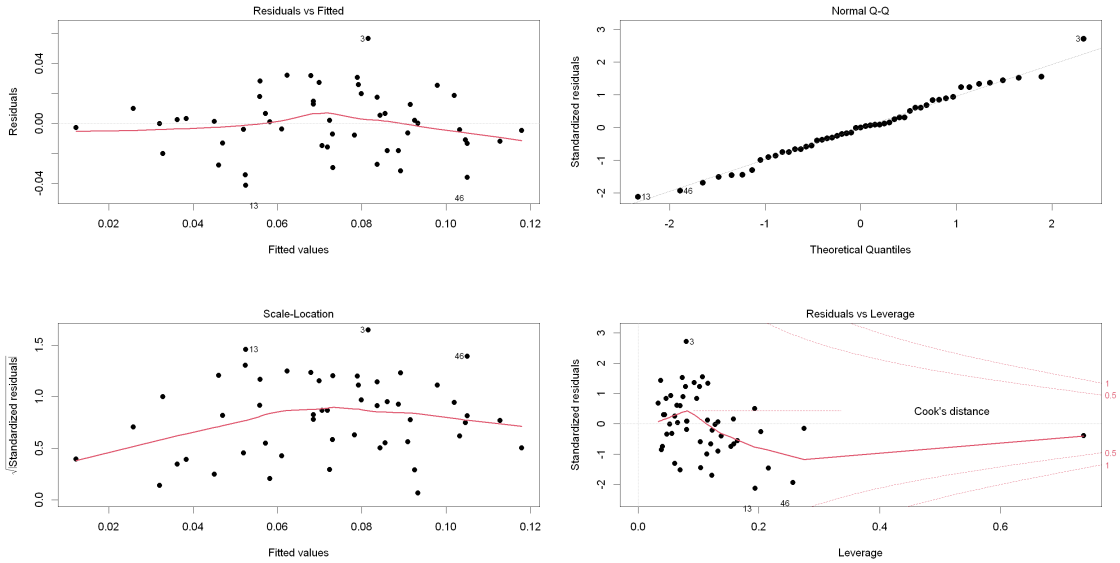
We use a combination of visual inspections and formal statistical tests to check on the 6 Classical Linear Model (CLM) assumptions:

- **MLR.1 Linearity in Parameters:** For model 2, we assume linearity in parameters by default. We have not transformed our dependent variable infection_rate and only log transformed log density to improve linearity with our dependent variable. Thus, MLR.1 holds.
- **MLR.2 Random Sampling:** We unfortunately cannot assume that MLR.2 holds as they may not be true random samples. Individuals who are sick are more likely to get tested while Individuals who may be sick, but are asymptomatic, may be afraid of getting testing due to social stigma. However, the p-value of 0.598 from the Durbin-Watson below indicates that there is no autocorrelation in the dataset. Thus, we can still perform linear modeling because we assume the sample is still representative of the underlying population.
- **MLR.3 No Perfect Collinearity:** We used the Variance Inflation Score (VIF) to determine if there are symptoms of multicollinearity among our independent variables (see VIF score below), which is not easily identifiable through the diagnostic plots.
  - **1. Residuals vs. Fitted:** The plot reveals a slight dip, but not enough to suggest that the assumption of zero conditional mean of errors have been violated in this model. Thus, **MLR.4** holds.
  - **2. Normal Q-Q:** This plot shows if residuals are normally distributed if it follows a relatively straight line. From the plot below, the bulk of the error terms seem to follow the straight line which suggests a fairly normal distribution. Additionally, the Shapiro-Wilk normality test returned a p-value of 0.8153 (see test below plots), further supporting the evidence that the normality of errors/residuals assumption holds (**MLR.6**)
  - **3. Scale-Location:** This plot shows if residuals are spread equally along the ranges of explanatory variables. From the plot below, the red line has a slight bend, but the residuals are relatively equally (random) spread. Thus, we believe the homoscedasticity

(homogeneity of residuals' variance) assumption holds (**MLR.5**). This claim is also supported by the p-value of 0.7066 from the Breusch-Pagan test (see test below plots).

- **4. Residuals vs. Leverage:** This plot helps us find influential or high leverage points if any. An influential value is a value can alter the results of the regression analysis and is associated with a large residual. In our specific plot below, all of the values are within the Cook line.

[37]:
```
par(mfrow = c(2, 2), cex = 1.5)
plot(model2, cex.caption = 1.5, cex.axis = 1, cex.lab = 1, pch = 20, cex = 1.5,␣
 ↪lwd = 3)
```



Shapiro-Wilk normality test (MLR.6): If p-value > 0.05, then it implies that the distribution of the data are not significantly different from the normal distribution.

[38]:
```
shapiro.test(model2$residuals)
```

```
Shapiro-Wilk normality test

data:  model2$residuals
W = 0.98625, p-value = 0.8153
```

Breusch-Pagan test (MLR.5): If p-value < 0.05, then it implies the null hypothesis of homoscedasticity should be rejected.

[39]:
```
bptest(model2)
```

```
studentized Breusch-Pagan test
```

```
data:  model2
BP = 2.9571, df = 5, p-value = 0.7066
```

Variance Inflation Factor (VIF) scores (MLR.3): Quantifies the extent of correlation between one predictor and the other predictor variables in a regression model; the higher the value, the greater the correlation of the variable with other variables. Values of more than 5 are sometimes regarded as problematic.

From the output below, all of our predictor variables are well below the 5 threshold, so we can claim that there is no perfect collinearity (**MLR.3**) in model 2.

[ ]:
```
car::vif(model2)
```

Variable	VIF	x < 5?
sip_start_end	1.20	TRUE
log(pop_dens)	1.42	TRUE
perc_atrisk_covid	1.46	TRUE
age_26to34	2.00	TRUE
age_55to64	1.71	TRUE

Durbin-Watson test (MLR.2): The Durbin Watson (DW) statistic is a test for autocorrelation in the residuals and we can use the output to help assess our random sampling assumption. Specifically, a p-value greater than .05 means that we cannot reject the null hypothesis that there is no autocorrelation among the residuals.

[61]:
```
durbinWatsonTest(model2)
```

```
 lag Autocorrelation D-W Statistic p-value
   1      0.04986469      1.852434   0.598
 Alternative hypothesis: rho != 0
```

Based on our review of the CLM assumptions, we can say that model 2 is an unbiased estimator of infection rate.

3.4 Model 3

3.4.1 Independent Variable Selection

Our 3rd model includes all variables in model 2 in addition to the additional variables listed below. We've excluded certain variables because they were highly correlated with other independent variables in the model.

- **age_35to54:** This age interval had a negative correlation with infection rate and represent percent of population in this age interval
- **totalTestResults_100k:** Represents the total number of Covid-19 tests administered and returned. Increasing testing capacity could increase case count due to asymptomatic carriers

of the disease

- **CaseRate_100k:** Measures the total number of positive cases per 100,000 people. This definition is different than infection rate because it is dividing the total positive case count by the state's 2018 population instead of total test results. It's also highly correlated with totalTestResults_100k, but we opted for CaseRate_100k instead due to its proximity to infection rate
- **CasesInLast7Days_100k:** A sliding window variable that counts the number of positive cases in the last 7 days per 100,000 people
- **soe_to_sip:** Measures the duration of non-essential businesses closured
- **party16:** This is the winning political party to the 2016 Presidential Election for each state
- **party2016_votepercent:** Represents the percentage of total popular vote the winning party received
- **biz_close_open:** Measures the duration of non-essential businesses closure
- **soe_face_mask:** Number of days between the state of emergency was declared and a face mask mandate for public facing employees was ordered
- **perc_under_fed_povline18:** Proportion of the population under the Federal poverty line based on 2018 readings

We excluded age_65plus, age_0to18 and age_19-to25 from the all-inclusive model because all 3 variables were highly correlated with the age_55to64 variable. In addition, we also excluded total cases, total death and Death_100k because we believe their effects have been partially absorbed by CaseRate_100k and CasesInLast7Days_100k. For death related metrics, their statistics have already been incorporated into the positive case count, since one must be positive for Covid-19 in order to pass away from the disease. We also opted for totalTestResults_100k instead of totalTestResults because the former takes into consideration population density and is consistent with other case related metrics (i.e. CaseRate_100k and CasesInLast7Days_100k).

3.4.2 Model Execution and Interpretation

Based on the output below, the bulk of the increase in the adjusted R^2 (from 0.5255 in model 2 to 0.7575 in model 3) comes from the CaseRate_100k and totalTestResults_100k variables. Besides that, 4 out of the 5 independent variables that were statistically and practically significant in model 2 were also significant in both manners in model 3. However, age_55to64 is no longer statistically significant. In addition, any increase in case count, whether it's case rate per 100k or case increases in the last 7 days, will always impact infection rate, since the numerator is always total case count. If a state ramps up testing, but positive case count is low, then we would naturally expect a lower infection rate. Moreover, our research question isn't whether increase in positive cases will impact infection rate because that's also impacted by testing capacity of each state. Rather, we want to understand which policy, demographic and state characteristic variables are significant factors to Covid-19 infection rate. Because population density (pop_dens), the duration of shelter-in-place (sip_start_end), individuals ageed 26 to 34 (age_26to34) and percent of the population who are at risk of serious illness from Covid-19 (perc_atrisk_covid) have consistently proven to be significant factors in previous models, we believe model 2 to be the most robust with respect to our research question.

Interpretation: Population density, perc_atrisk_covid, age_26to34, and duration of shelter-in-place have not changed from model 2. For CaseRate_100k, we can say that for every 1 unit increase in positive cases per 100,000 people, we can expect infection rate to increase by 0.00003

percentage points holding all other variables constant. Additionally, for every 1 unit increase in totalTestResults_100k, we can expect the infection rate to also decrease by 0.000004 percentage points holding all other variables constant. totalTestResults_100k is interesting because it is saying that the more tests are administered, the lower the infection rate. This make sense because as the number of tests approaches infinity, infection rate will reach 0, since there are only so many people in the world that we can test (assuming no resource or capacity constraint, which is unlikely). In other words, people are finite while testing kits can be infinite, and if we kept testing everyone over and over again, infection rate will eventually approach 0, since we are increasing our definition of infection rate's denominator at a faster rate than positive cases.

[41]:
```
model3 = lm(infection_rate ~ sip_start_end + log(pop_dens) + perc_atrisk_covid␣
↪+ age_26to34 + age_55to64 + age_35to54 + totalTestResults_100k +␣
↪CaseRate_100k + CasesInLast7Days_100k + soe_to_sip + party16 +␣
↪party2016_votepercent +
            biz_close_open + soe_face_mask + perc_under_fed_povline18 +␣
↪wunemp_insure_max, data = df_covid2)
summary(model3)
```

```
Call:
lm(formula = infection_rate ~ sip_start_end + log(pop_dens) +
    perc_atrisk_covid + age_26to34 + age_55to64 + age_35to54 +
    totalTestResults_100k + CaseRate_100k + CasesInLast7Days_100k +
    soe_to_sip + party16 + party2016_votepercent + biz_close_open +
    soe_face_mask + perc_under_fed_povline18 + wunemp_insure_max,
    data = df_covid2)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0301204 -0.0102749  0.0005311  0.0093739  0.0238270

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               3.308e-01  1.118e-01   2.960 0.005577 **
sip_start_end            -4.013e-04  1.188e-04  -3.379 0.001839 **
log(pop_dens)             1.037e-02  2.657e-03   3.905 0.000425 ***
perc_atrisk_covid        -4.607e-03  1.523e-03  -3.024 0.004718 **
age_26to34               -7.917e-01  2.932e-01  -2.700 0.010725 *
age_55to64                4.881e-02  3.828e-01   0.127 0.899309
age_35to54               -2.823e-01  3.216e-01  -0.878 0.386217
totalTestResults_100k    -3.910e-06  9.042e-07  -4.324 0.000127 ***
CaseRate_100k             2.977e-05  5.441e-06   5.472 4.18e-06 ***
CasesInLast7Days_100k     6.893e-05  4.514e-05   1.527 0.136017
soe_to_sip                2.556e-04  4.889e-04   0.523 0.604460
party16                   6.167e-03  8.102e-03   0.761 0.451822
party2016_votepercent     6.659e-03  4.597e-02   0.145 0.885669
biz_close_open            3.856e-04  2.246e-04   1.717 0.095064 .
soe_face_mask             1.211e-04  1.096e-04   1.105 0.276734
perc_under_fed_povline18  2.996e-03  1.572e-03   1.906 0.065137 .
```

```
wunemp_insure_max        -1.595e-05  2.330e-05  -0.685 0.498251
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01552 on 34 degrees of freedom
Multiple R-squared:  0.8351,Adjusted R-squared:  0.7575
F-statistic: 10.76 on 16 and 34 DF,  p-value: 5.236e-09
```

While the standard errors indicate statistical significance the 6 starred variables above, the heteroskedastic-robust errors tells a completely different story, one in which none of the variables are statistically significant

[42]:
```
coeftest(model3, vcov = vcovHC, level = 0.05)
```

```
t test of coefficients:

                           Estimate  Std. Error t value Pr(>|t|)
(Intercept)               3.3077e-01  2.4766e-01  1.3356   0.1906
sip_start_end            -4.0133e-04  3.0619e-04 -1.3107   0.1987
log(pop_dens)             1.0374e-02  1.2211e-02  0.8496   0.4015
perc_atrisk_covid        -4.6074e-03  3.6318e-03 -1.2686   0.2132
age_26to34               -7.9175e-01  8.1939e-01 -0.9663   0.3407
age_55to64                4.8806e-02  5.7312e-01  0.0852   0.9326
age_35to54               -2.8226e-01  5.2421e-01 -0.5385   0.5938
totalTestResults_100k    -3.9101e-06  2.3802e-06 -1.6428   0.1096
CaseRate_100k             2.9774e-05  5.5123e-05  0.5401   0.5926
CasesInLast7Days_100k     6.8927e-05  1.0467e-04  0.6585   0.5147
soe_to_sip                2.5561e-04  6.7489e-04  0.3787   0.7072
party16                   6.1670e-03  1.3270e-02  0.4647   0.6451
party2016_votepercent     6.6593e-03  7.1024e-02  0.0938   0.9258
biz_close_open            3.8563e-04  3.9641e-04  0.9728   0.3375
soe_face_mask             1.2112e-04  1.3982e-04  0.8663   0.3924
perc_under_fed_povline18  2.9959e-03  2.7393e-03  1.0937   0.2818
wunemp_insure_max        -1.5949e-05  2.7582e-05 -0.5783   0.5669
```

3.4.3 Model Diagnostics

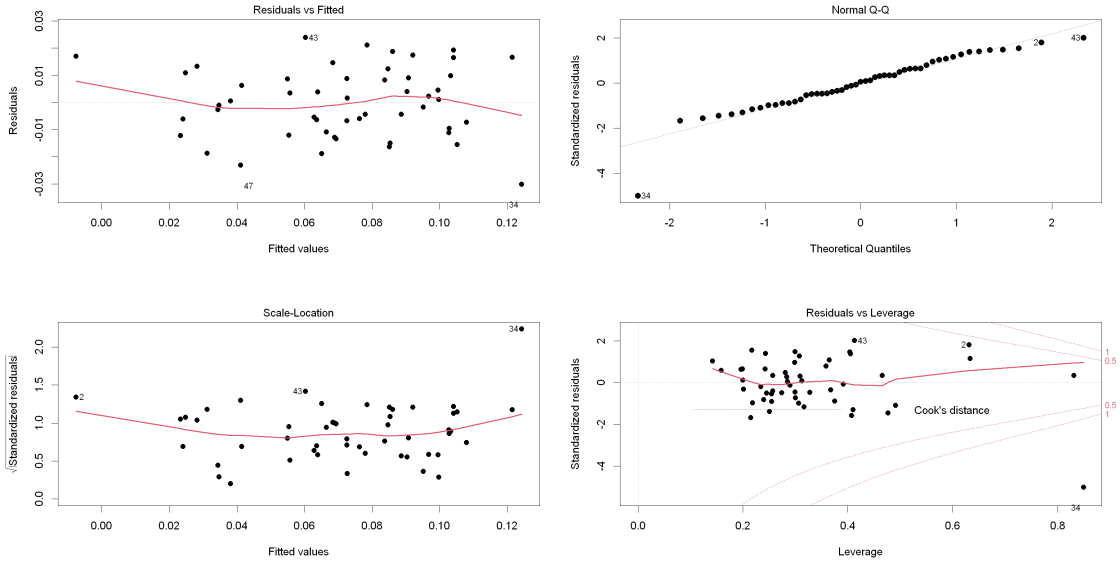
We use a combination of visual inspections and formal statistical tests to check on the 6 Classical Linear Model (CLM) assumptions:

- **MLR.1 Linearity in Parameters:** For model 3, we assume linearity in parameters by default. We have not transformed our dependent variable infection_rate and only log transformed log density to improve linearity with our dependent variable. Thus, MLR.1 holds.
- **MLR.2 Random Sampling:** We unfortunately cannot assume that MLR.2 holds as they may not be true random samples. Individuals who are sick are more likely to get tested while

Individuals who may be sick, but are asymptomatic, may be afraid of getting testing due to social stigma. Additionally, the p-value of 0.016 from the Durbin-Watson below indicates that there is evidence of autocorrelation in the residuals. While this is not true random sampling, we believe we can still perform linear modeling because we assume this sample is representative of the underlying population.

- **MLR.3 No Perfect Collinearity:** We used the Variance Inflation Score (VIF) to determine if there are symptoms of multicollinearity among our independent variables (see VIF score below), which is not easily identifiable through the diagnostic plots.
  - **1. Residuals vs. Fitted:** The plot reveals a slight bend, but not enough for us to reject the assumption of zero conditional mean of errors is violated. Thus we assume **MLR.4** holds
  - **2. Normal Q-Q:** This plot shows if residuals are normally distributed if it follows a relatively straight line. From the plot below, the bulk of the error terms seem to follow the straight line which suggests a fairly normal distribution. Additionally, the Shapiro-Wilk normality test returned a p-value of 0.5955 (see test below plots), further supporting the evidence that the normality of errors/residuals assumption holds (**MLR.6**)
  - **3. Scale-Location:** This plot shows if residuals are spread equally along the ranges of explanatory variables. From the plot below, while the red line is horizontal, we do see a pattern of residuals concentrated in the center, which means they are not randomly spread. Thus, we believe the homoscedasticity (homogeneity of residuals' variance) assumption does not holds (**MLR.5**) for model 3. This claim is also supported by the p-value of 0.02703 from the Breusch-Pagan test (see test below plots).
  - **4. Residuals vs. Leverage:** This plot helps us find influential or high leverage points if any. An influential value is a value can alter the results of the regression analysis and is associated with a large residual. In our specific plot below, observation 34 (North Carolina) is above the 1 Cook line. Upon further investigation, North Carolina did not order public facing employees to wear face covers until 6/26, 108 days after a state of emergency was declared. Because this represent specific choices made by legislators in North Carolina, we cannot simply remove it without practical justification.

[43]:
```
par(mfrow = c(2, 2), cex = 1.5)
plot(model3, cex.caption = 1.5, cex.axis = 1, cex.lab = 1, pch = 20, cex = 1.5,␣
↪lwd = 3)
```



Shapiro-Wilk normality test (MLR.6): If p-value > 0.05, then it implies that the distribution of the data are not significantly different from the normal distribution.

[44]:
```
shapiro.test(model3$residuals)
```

```
Shapiro-Wilk normality test

data:  model3$residuals
W = 0.98131, p-value = 0.5955
```

Breusch-Pagan test (MLR.5): If p-value < 0.05, then it implies the null hypothesis of homoscedasticity should be rejected.

[45]:
```
bptest(model3)
```

```
studentized Breusch-Pagan test

data:  model3
BP = 28.566, df = 16, p-value = 0.02703
```

Variance Inflation Factor (VIF) scores (MLR.3): Quantifies the extent of correlation between one predictor and the other predictor variables in a regression model; the higher the value, the greater the correlation of the variable with other variables. Values of more than 5 are sometimes regarded as problematic.

From the output below, most of our predictor variables are well below the 5 threshold, except

for perc_atrisk_covid. Although its VIF score is above 5, we cannot say that there is perfect collinearity. Thus, we believe model 3 still satisfies the **MLR.3.**

[ ]:
```
car::vif(model3)
```

Variable	VIF	x < 5?
sip_start_end	2.67	TRUE
log(pop_dens)	3.26	TRUE
perc_atrisk_covid	**6.56**	**FALSE**
age_26to34	4.04	TRUE
age_55to64	4.65	TRUE
age_35to54	2.12	TRUE
totalTestResults_100k	2.63	TRUE
CaseRate_100k	2.55	TRUE
CasesInLast7Days_100k	2.92	TRUE
soe_to_sip	2.27	TRUE
party16	3.37	TRUE
party2016_votepercent	2.62	TRUE
biz_close_open	2.04	TRUE
soe_face_mask	1.52	TRUE
perc_under_fed_povline18	4.14	TRUE
wunemp_insure_max	2.15	TRUE

Durbin-Watson test (MLR.2): The Durbin Watson (DW) statistic is a test for autocorrelation in the residuals and we can use the output to help assess our random sampling assumption. Specifically, a p-value greater than .05 means that we cannot reject the null hypothesis that there is no autocorrelation among the residuals.

[62]:
```
durbinWatsonTest(model3)
```

```
 lag Autocorrelation D-W Statistic p-value
   1       0.3006965       1.34551   0.016
 Alternative hypothesis: rho != 0
```

Based on our review of the CLM assumptions, we do not believe that model 3 is an unbiased estimator of infection rate because it violates MLR.5. In addition, 2 of the statistically significant variables (based on normal standard errors) in the model do not provide practical significance relative to our research question. Moreover, if we replaced the each predictor variable's normal standard errors with heteroskedastic-robust errors, none of the independent variables in model 3 are statistically significant, which provides further evidence that our more parsimonious model 2 is robust. Nonetheless, we do believe omitted variables play a major role in the lack of explanatory power, since policies are only meant to curtail activities and behaviors that increase the risk of exposure and contraction if it enforced and carries a heavy penalty.

4 Regression Table

[47]:
```
# Replace regular Standard Errors with the heteroscedasticity-robust Standard␣
↪Errors
se.model1 <- sqrt(diag(vcovHC(model1)))
se.model2 <- sqrt(diag(vcovHC(model2)))
se.model3 <- sqrt(diag(vcovHC(model3)))
```

[48]:
```
stargazer(model1, model2, model3, title = "Regression Model Comparison", type =␣
↪"text", report = "vc*s", omit.stat = "f", se = list(se.model1, se.model2, se.
↪model3),
          star.cutoffs = c(0.05, 0.01, 0.001), font.size = "normalsize", column.
↪sep.width = "2pt", align = TRUE, column.labels = c("Model 1", "Model 2",␣
↪"Model 3"))
```

```
Regression Model Comparison
=============================================================================
                                      Dependent variable:
                          ---------------------------------------------
                                          infection_rate
                             Model 1          Model 2          Model 3
                               (1)              (2)              (3)
-----------------------------------------------------------------------------
biz_close_open                0.001                             0.0004
                             (0.0004)                          (0.0004)

soe_face_mask                                                   0.0001
                                                               (0.0001)

perc_under_fed_povline18                                         0.003
                                                                (0.003)

wunemp_insure_max                                              -0.00002
                                                               (0.00003)

sip_start_end               -0.0004*         -0.0004**          -0.0004
                             (0.0002)         (0.0001)          (0.0003)

soe_to_sip                    0.001*                            0.0003
                              (0.001)                           (0.001)

party16                                                          0.006
                                                                (0.013)

party2016_votepercent                                            0.007
                                                                (0.071)
```

```
log(pop_dens)                                0.017***          0.010
                                             (0.002)          (0.012)

perc_atrisk_covid                            -0.002           -0.005
                                             (0.001)          (0.004)

age_26to34                                   -1.100***        -0.792
                                             (0.275)          (0.819)

age_55to64                                   -0.969*           0.049
                                             (0.446)          (0.573)

age_35to54                                                    -0.282
                                                              (0.524)

totalTestResults_100k                                        -0.00000
                                                             (0.00000)

CaseRate_100k                                                 0.00003
                                                             (0.0001)

CasesInLast7Days_100k                                         0.0001
                                                             (0.0001)

Constant                          0.033      0.362***          0.331
                                 (0.022)     (0.098)          (0.248)

-------------------------------------------------------------------------
Observations                       51           51               51
R2                                0.191        0.573            0.835
Adjusted R2                       0.140        0.525            0.757
Residual Std. Error      0.029 (df = 47) 0.022 (df = 45) 0.016 (df = 34)
=========================================================================
Note:                                        *p<0.05; **p<0.01; ***p<0.001
```

In addition to the standard regression metrics, we also leverage Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) to evaluate the trade-off between goodness of fit and model complexity/parsimony. In other words, we not want a model with high explanatory power, but also a parsimonious one.

The **AIC** is a relative measure of model parsimony. A lower score indicates a more parsimonious model relative to a model fit with a higher AIC. Thus, we want the model with the lowest AIC. While model 3 has the lowest AIC score, we argue that because none of its predicator variables are statistically significant when we use the heteroskedastic-robust errors and it violates MLR.5, we still believe model 2 strikes the best balance between goodness of fit and parsimony.

[49]:
```
AIC(model1, model2a, model2, model2b, model3)
```

		df	AIC
		<dbl>	<dbl>
A data.frame: 5 × 2	model1	5	-209.7528
	model2a	9	-221.0642
	model2	7	-238.3088
	model2b	6	-235.2060
	model3	18	-264.8371

The **BIC** is similar to the AIC, but has a larger penalty, and a lower score indicates a more parsimonious model relative to a model fit with a higher BIC. From the output below, the difference between model 2 and model 3 is much smaller than AIC. For the same reason explained in the previous section, we still believe model 2 is a more robust and consistent model than model 3.

[50]:
```
BIC(model1, model2a, model2, model2b, model3)
```

		df	BIC
		<dbl>	<dbl>
A data.frame: 5 × 2	model1	5	-200.0937
	model2a	9	-203.6778
	model2	7	-224.7861
	model2b	6	-223.6151
	model3	18	-230.0642

5 Omitted Variables Discussion

Below are the top omitted variables that we believe are important along with how their absence may affect our results:

- **Adherence to Policy:** Most states have not enforced the violation of shelter in place policies. This is contrary to other countries, such as China, where a more draconian set of measures were put into effect to quickly stop the spread of the disease. Referencing other countries such as China and Italy who prosecuted individuals who broke such policies, they were able to contain the infection rate in a matter of weeks. A proxy measure could be used to operationalize such metric is the change in user's geo-location; if someone moved around more frequently than the city's norm/benchmark, then we can get a directional indicator of how tightly that individual is adhering to social distancing policies.
  - **Impact of bias:** We believe that if adherence increases, infection rate will decrease. Additionally, increased adherence is also likely to shorten the duration of restrictive policies such as shelter-in-place. Thus, we believe that the coefficient of sip_start_end will be scaled toward 0 and becoming less negative (losing statical significance).
- **General Face Covering Mandate:** Although the dataset contained a variable on when face coverings for public-facing employees were required, it does not include any data on whether the state issued face covering requirement for all residents. There is emerging evidence for the effectiveness of cloth face coverings Referencing Asian countries that have effectively suppressed the spread of Covid-19, face coverings were part of daily life and much more accepted in their culture than it is in the United States.
  - **Impact of bias:** We believe that if general face covering was mandatory for everyone when venturing outside, infection rate will decrease. Additionally, the mandate is also likely to shorten the duration of restrictive policies such as shelter-in-place. Thus, we

believe that the coefficient of sip_start_end will be scaled toward 0 and becoming less negative (losing statical significance).

- **Population by Race:** There is a growing body of research that indicates minority groups are contracting and dying from Covid-19 at a disproportionate rate than whites. An article published by the CDC provides evidence that "Long-standing systemic health and social inequities have put some members of racial and ethnic minority groups at increased risk of getting Covid-19 and/or experiencing severe illness, regardless of age." We hypothesize that states with a higher proportion of minorities and people of color relative to whites would increase both infection and fatality rate. Due to the growing body of evidence that's emerging from public health institutions and academia, we believe racial variables would have a moderate to high effect size on infection rate and fatality rate.
  - **Impact of bias:** We believe the proportion of racial minorities in a given state is positively correlated with infection rate. Additionally, it's also likely to increase the percent of individuals of becoming seriously ill. Thus, we believe that the coefficient of perc_atrisk_covid will be scaled away from 0 and becoming more positive (gaining statical significance).
- **Public Transit Usage:** According to this article, the use of public transits are associated with increased Covid-19 infection and fatality rate. This makes intuitive sense, since buses, subways and light rails are confined spaces loaded with numerous people that are traveling to and from different locations. This is especially problematic because not only do these transits increase the likelihood of transmission within the confines of the modality, it also enables the disease to go mobile, since infected individuals are also traveling, potentially infecting others along the way. For these reasons, we believe that a decreased use of public transits can decrease the transmission and infection of Covid-19. We also hypothesize that this variable would elicit a medium to high effect size on infection rate conditioned on population density.
  - Impact of bias: We believe increased public transit usage is positively correlated with infection rate. Additionally, it's also likely to increase the percent of individuals of becoming seriously ill. Thus, we believe that the coefficient of perc_atrisk_covid will be scaled away from 0 and becoming more positive (gaining statical significance).
- **County Level Re-Opening Dates:** We believe county level re-opening dates to be important data with respect to infection rate because certain states such as California have partially delegated the re-opening decision to each county conditioned that they meet certain metrics. County level information would provide additional samples that can be used for modeling from a goodness of fit and explanatory power perspective. More importantly, county level data could provide policy makers a more precise set of tools to create tailored approaches to reduce the spread of Covid-19; a one size fits all approach does not consider the potential differences in demographics, population, racial diversity, and socio-economic statuses of the state (example of California).
  - Impact of bias: In general, we believe that the sooner the state re-opens since restrictive policies were put into effect, the higher the infection rate. Additionally, it's also positively correlated with the duration of shelter-in-place order. Thus, we believe that the coefficient of sip_start_end will be scaled toward 0 and become less negative (losing statical significance).
- **Political Leanings:** The politicization by the Republican and Democratic parties could have a large effect on the number of individuals willing to follow state mandated guidelines. Another potential variable in this search could be state legislatures majority political party, as they would be the governing body that had political power to enact policies during the start of the pandemic. We believe this omitted variable may not only be of high correlation

between citizens that adhere to policy but may actually be causal as the pandemic has been highly politicised. Further more, data shows that Republican leaning states are less likely to adopt Covid aware policies, which could account for the differences in policies that have been noted by each state (fivethirtyeight). We believe this omitted variable might shine a light on some hidden patterns in the data and allow our models to better predict where life saving measures have been subverted.

- Impact of bias: In general, we believe that the proportion of eligible voters that self-identify as Republicans is positively correlated with the infection. Additionally, it's likely to negatively correlate with the duration of shelter-in-place policy (or even the absence of such order). Thus, we believe that the coefficient of sip_start_end will be scaled away from 0 and become more negative (gaining statical significance).

6 Conclusion

Based on our analysis and the different variants of regression models that were built, we conclude that state characteristic coupled with demographic factors were more influential than legislative policies with respect to Covid-19 infection rate. While we do not deny that responses such as shelter-in-place orders or non-essential business closures have discouraged and significantly decreased social activities and gatherings, our model provides minimal evidence on their association with infection rate. In addition to policies, we also assessed the timeliness of policy declaration and the duration of each policy. Our initial hypothesis was that the sooner a policy was enacted after the state of emergency was declared, the lower the infection rate. Similarly, the longer a social distancing order was in put into effect, the lower we would expect the infection rate to be. However, we were only partially correct, which is reflected in the duration of shelter-in-place orders.

In conclusion, while our and analysis and model provides a starting point to identify policy, social and economic factors that are associated with Covid-19 infection rate, we believe there are a host of other variables and data that are needed to create a model of practical value and higher explanatory power. The global pandemic is only 7 months deep and we believe the data currently being collected will fuel further research not only in disease prevention, but data-driven legislations, responses and overall emergency preparedness to better manage the next outbreak.